OPEN ACCESS

University of BRISTOL

This electronic thesis or dissertation has been downloaded from Explore Bristol Research, http://research-information.bristol.ac.uk

Author:
Stevens, Emily

Title:
Investigating the genetic basis for bacterial toxin production using functional genomics

General rights
Access to the thesis is subject to the Creative Commons Attribution - NonCommercial-No Derivatives 4.0 International Public License. A copy of this may be found at https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode This license sets out your rights and the restrictions that apply to your access to the thesis so it is important you read this before proceeding.

Take down policy
Some pages of this thesis may have been removed for copyright restrictions prior to having it been deposited in Explore Bristol Research. However, if you have discovered material within the thesis that you consider to be unlawful e.g. breaches of copyright (either yours or that of a third party) or any other law, including but not limited to those relating to patent, trademark, confidentiality, data protection, obscenity, defamation, libel, then please contact collections-metadata@bristol.ac.uk and include the following information in your message:

• Your contact details
• Bibliographic details for the item, including a URL
• An outline nature of the complaint

Your claim will be investigated and, where appropriate, the item in question will be removed from public view as soon as possible.

Investigating the genetic basis for bacterial toxin production using functional genomics

Emily Jane Stevens

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Faculty of Biomedical Sciences, School of Cellular and Molecular Medicine.

September, 2019

Word count: 56,167

Abstract

The genetic networks that regulate bacterial phenotypes are highly complex and incompletely understood in many cases. One such example of this is virulence, which is dependent on a variety of contributors including genetic, environmental and host factors. If we are to continue to develop new treatments and preventative measures against bacterial disease, it will become increasingly important to understand how different pathogens cause disease, and genome sequencing information can be of considerable use in doing so. A key aspect of bacterial virulence is toxicity, which is the production of proteins toxic to host cells that actively damage them to the benefit of the bacteria. This work uses functional genomics to further elucidate the genetic basis for toxicity in two medically important pathogens; *Staphylococcus aureus* and *Streptococcus pneumoniae*. Following identification of the *cyoE* locus as a novel effector of toxicity in *S. aureus* using a genome-wide association study (GWAS), this gene is shown to be involved in processing haem for incorporation into the electron transport chain during aerobic respiration. In its absence *S. aureus* respires as if growing in microaerobic conditions and down-regulates toxin production. The functional genomics approach is then extended further in *S. pneumoniae*, using three different GWAS methods in combination to aid identification of novel effectors of the toxic phenotype. From this, the SPN23F12470 gene is identified as a novel effector of toxicity in the pneumococcus, which encodes a UvrD-like protein with helicase and nuclease domains. Data presented here shows a physical interaction between this protein and regions of the genome either side of the pneumococcal toxin pneumolysin which may be interfering with expression of the toxin gene. Functional genomics is therefore shown to be a useful approach in the identification of novel effectors of toxicity in *S. aureus* and *S. pneumoniae*.

Acknowledgements

Many thanks to my supervisor Ruth Massey for her guidance and support, and to my co-supervisor Mario Recker for his support in running our GWAS analysis.

Many thanks also to Maisem Laabei for his continued advice and support, and to all members of the Massey lab past and present.

Thanks to all our collaborators who have helped with many aspects of this work, particular thanks go to:

Greg Somerville, Stewart Gardner and Jean van den Elsen for their assistance and insight into the role of the *cyoE* gene.

Stephen Bentley, Rebecca Gladstone and Nick Croucher for providing the collection of clinical *Streptococcus pneumoniae* PMEN-1 isolates and the associated SNP data which was used for our GWAS analysis.

Angela Nobbs, Mikaila Bandara, Howard Jenkinson, Tim van Opijnen and Derek Thibault for providing *S. pneumoniae* mutants and sharing their expertise regarding how to construct knockouts.

John Lees, Daniel Wilson and Sarah Earle for providing alternative GWAS analyses.

Mark Dillingham and Oliver Wilkinson for their help conducting electrophoretic mobility shift assays and their insights into the function of the UvrD-like protein.

Many thanks to the BBSRC SWBio DTP programme for funding this work (grant reference BB/M009122/1).

Thanks also to all the following people:

Tammy Hassel for her support whilst on placement at Microgenetics.

Eilis Bragginton and Victoria Knox for welcoming me so warmly into the lab at Bristol when we moved.

Katy Sutcliffe and Clara Montgomery for the many Wednesday evenings spent in the pub.

Tom Belcher for his support and encouragement, for bringing out the best in me and never letting me give up.

My wonderful family for their support and for always inspiring me to aim high.

Author’s declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: ..... DATE:.....

Publications

Stevens, E. et al., 2017. Cytolytic toxin production by *Staphylococcus aureus* is dependent upon the activity of the protoheme IX farnesyltransferase. *Scientific Reports*, 7(1). DOI: 10.1038/s41598-017-14110-8.

Contents

List of figures

List of tables

1. Introduction

Bacterial virulence is a highly complex and multi-faceted phenotype which encompasses the many aspects of a bacterium's ability to cause disease. Here, the term 'virulence' is used to refer to factors which can be broadly classified into three groups based on their functions in different stages of disease, namely: factors associated with colonisation and invasion of host tissues; factors involved in evasion of the host immune system; and toxic factors involved in actively damaging host cells.

The work presented in this thesis builds on previous research by our group to contribute to a better understanding of the genetic basis for a key aspect of bacterial virulence; toxicity. In previous work by Laabei et al. (2014; 2015), a genome-wide association study (GWAS) was conducted on large collections of clinical *Staphylococcus aureus* isolates which led to the identification of novel genetic loci that contribute to toxin production in this species. The work presented here follows on from this project with the detailed characterisation of one of these loci, extension of the approach to the bacterium *Streptococcus pneumoniae*, and finally with the characterisation of a novel toxicity-affecting locus identified in *S. pneumoniae*.

This approach, starting with whole genome sequence data and ending with the characterisation of a toxicity-affecting gene, is referred to as functional genomics. The results presented here demonstrate how this approach has been successfully used in two bacterial species to identify novel toxicity-affecting genetic loci, starting with characterisation of the GWAS-identified locus *cyoE* in *S. aureus*, which has not previously been reported to be involved in bacterial toxicity. In *S. pneumoniae*, the approach was extended to include a range of GWAS methods which were used in combination to pinpoint more accurately those loci which are associated with the toxic phenotype. This resulted in identification of the *uvrD*-like gene SPN23F12470, which showed a reduction in toxicity when knocked out, and is subsequently characterised to elucidate its effect on toxin production in *S. pneumoniae*.

Staphylococcus aureus and *Streptococcus pneumoniae* are medically important pathogens, both of which can survive as harmless commensals in the human

host, but which can also transition to opportunistic pathogens in certain circumstances. Both species occupy a similar niche within the host, living in the nasopharynx in their commensal state (Bogaert, de Groot & Hermans, 2004; Wertheim et al., 2005), and are known to compete for colonisation of this area (Park et al., 2008). There is a substantial difference in toxin production between the two species, with *Staphylococcus aureus* possessing fifteen known toxins employed in different types of disease, while *Streptococcus pneumoniae* has only one known toxin. However, the work presented here demonstrates that toxin production in both species relies on a complex network of genetic contributors outside of the well-characterised toxin genes, which is not yet fully understood.

1.1.0 From commensal to virulent pathogen

A characteristic feature of both *S. aureus* and *S. pneumoniae* is their ability to transition from commensal to opportunistic pathogen within the human host. *S. aureus* can be commonly found in the nasal passages of around 20-30% of the population (Gordon & Lowy, 2008) as a commensal, but can become the cause of opportunistic infections within the host when the immune system becomes compromised through illness, injury or surgery. Staphylococcal infections can range from mild skin and soft tissue infections, to serious invasive disease including pneumonia, meningitis and septicaemia. Likewise, *S. pneumoniae* is commonly carried in up to 60% of young children (Henriques-Normark & Normark, 2010) and 2-20% of adults (Regev-Yochay et al., 2003; Hussein et al., 2005; Henriques-Normark & Normark, 2010), but can cause a range of diseases including sinusitis, otitis media and the more invasive pneumonia, meningitis and septicaemia. Both can also be responsible for device-associated infections due to their ability to form biofilms on surgical implants, ventilators and catheters, for example (Rosenthal et al., 2006; Liu et al., 2012; Haque et al., 2018).

As commensals, these bacteria form part of the normal nasopharyngeal microflora which can act as barriers to other species of potentially pathogenic bacteria attempting to colonise the host. A stable balance is maintained through interactions between commensal bacteria, viruses and the host immune system, and its constituents vary greatly between individuals (Shak et al., 2013). Selective pressures such as the use of antibiotics and vaccines can also heavily influence the make-up of this ecosystem (Vergison, 2008). Competition for colonisation of this niche takes place predominantly between species of the families Moraxellaceae, Streptococcaceae, Corynebacteriaceae, Pasteurellaceae and

Staphylococcaceae (Shak et al., 2013). A number of studies have revealed an inverse correlation between carriage of *S. pneumoniae* and *S. aureus* (Chien et al., 2013; Regev-Yochay et al., 2004; Bogaert et al., 2004), with the pneumococcus inhibiting growth of staphylococci through production of hydrogen peroxide, to which *S. aureus* are sensitive (Regev-Yochay et al., 2006).

In a healthy host, these interactions usually prevent one bacterial species dominating the ecosystem and invading host tissues. On occasion, normally commensal bacteria cause infection in an otherwise healthy host, however, it is immunocompromised patients and people at the extremes of age who are most at risk of infection by species such as *S. aureus* and *S. pneumoniae*. In these cases, depression of the immune system can disturb the fine balance of the commensal bacterial community. As a result, one species outcompetes its rivals for space and nutrients within the ecosystem and becomes the dominant population. This species proliferates, taking advantage of a weakened host immune system, and as nutrients become increasingly scarce it invades host tissues, damaging host cells and consequently causing the symptoms of disease.

At this point, the invading species has transitioned from commensal to opportunistic pathogen and research has shown that in around 80% of cases of staphylococcal bacteraemia, the infecting strain in fact originated from the anterior nares of the patient (von Eiff et al., 2001). This transition requires an extensive array of virulence factors, including adhesins for initial colonisation of the host, immune evasion proteins for proliferation and persistence throughout infection, and toxins to destroy host cells for nutrient acquisition. This army of virulence factors available to breach host defences and establish infection is a key factor that differentiates obligate commensals from opportunistic pathogens (Casadevall & Pirofski, 1999).

Toxins in particular contribute greatly to the presentation of symptoms during infectious disease. For example, action of the toxin pneumolysin in the upper respiratory tract of pneumonia patients is responsible for the strong inflammatory response in the lung and the resulting fever and cough which are symptomatic of this disease (Boulnois, 1992). Increasingly, the role of toxins in different stages of pathogenesis is being highlighted, with recent research demonstrating that toxins are important throughout disease development, from colonisation through to presentation of symptoms (Rudkin et al., 2017). The biology of staphylococcal

and pneumococcal toxins is discussed in more detail in sections 1.3.2.3 and 1.4.2.4 respectively.

With ever-increasing levels of antibiotic resistance among bacterial pathogens, it is becoming increasingly important to understand exactly how species such as *S. aureus* and *S. pneumoniae* cause disease, such that we can work towards development of better prevention and treatment of disease. This encompasses a variety of factors which contribute to the onset of disease in a patient, from social factors such as hygiene procedures in hospitals, to individual factors such as underlying co-morbidities which may predispose someone to infection, and to the virulence potential of the infecting pathogen. This project focuses on the virulence potential of infecting pathogens, looking specifically to elucidate the genetic factors which contribute to toxin production and secretion in two prominent causes of opportunistic infection.

1.2.0 A genomic approach to studying virulence

The genetic basis for bacterial virulence is highly complex, and in an era where genome sequencing information is being obtained quickly and accurately, and used to track disease outbreaks and genetic changes in pathogens in real-time (Gardy et al., 2015; Quick et al., 2016), there is an increasing need to work towards a more complete understanding of the genetic basis of virulence.

Simply knowing the genome sequence of an isolate is of little use in the context of disease treatment unless we can fully understand the function of genes which contribute to the pathogenic phenotype. This point is highlighted by Viney (2014), where the importance is stressed of understanding individual gene function in an era when it is becoming increasingly common to compare similarities between genomes instead of focussing on the function of specific genes. If whole genome sequencing is to be used in the future to improve disease treatment, it is necessary to study individual gene function in relation to pathogenesis so that pathogenic genomes can be sequenced and interpreted in a way which will be clinically relevant. In particular, it is hoped that through an improved ability to accurately interpret genome sequencing information, it will one day be possible to diagnose and treat infections on an individual and personal basis, such that the most appropriate and effective treatment can be given as early as possible in the course of an infection.

This work takes a small step towards this aim by characterising the individual function of two genes which contribute to bacterial disease, one in *S. aureus* and one in *S. pneumoniae*. In doing so, this contributes to the long-term goal of being able to interpret genome sequencing information on a personalised basis to determine the potential severity of disease caused by a patient's individual bacterial isolate.

A recent genome-wide association study (GWAS) by Tunjungputri et al. (2017) took a further step towards this goal, demonstrating how sequencing data can be used to identify novel associations between genotype and phenotype in the context of bacterial disease. Specifically, the authors identified that presence of the *pblB* gene on the pneumococcal chromosome correlated with increased mortality in patients suffering from invasive pneumococcal disease, and that this was mediated by an increase in activation of platelets. Platelets have been shown to be important in patients suffering from sepsis, in that excessive platelet activation causes over-stimulation of the innate immune system, subsequently leading to organ failure and mortality. Identification of strains harbouring this gene during the course of an infection would give clinicians some indication of the potential severity of invasive pneumococcal disease which a patient may develop, and allow them to tailor treatment accordingly to the needs of the individual.

The need to determine the most appropriate treatment for infection is becoming ever more pressing considering the increased prevalence of antibiotic resistance observed in numerous bacterial pathogens over the past few decades (Gross, 2013; Levy, 1998; Neu, 1992); the rise of methicillin resistance in *Staphylococcus aureus* being one of the best examples of this. Investment into early-stage research for development of novel antibiotics has improved over recent years, however, investment into clinical development and commercialisation of novel antimicrobial compounds is lacking, with very few antibiotics currently in the later stages of clinical trials, and several large pharmaceutical companies terminating their antibiotic development programmes (Blaskovich et al., 2017; Simpkin et al., 2017). This lack of funding for the later stages of novel antibiotic development creates a barrier to new antibiotics entering the pharmaceutical market, exacerbating the long-term threat posed by antibiotic-resistant infections.

Given these challenges, it is essential that we work towards a better understanding of bacterial virulence, and in turn that this knowledge is used to

find alternative ways of treating the wide range of diseases caused by pathogens such as *S. aureus*. Potential alternative treatments include trying to boost host immunity to naturally fight off infection (Finlay & Hancock, 2004; Kyme et al., 2012), identifying potential new drug targets (Lee et al., 2009; Payne et al., 2007; Gotoh et al., 2010), and ensuring that the antibiotic treatments that are still effective are used efficiently and appropriately to minimise the further spread of resistance, and maintain a reserve of these antibiotics for last resort treatment.

Our lab group has previously studied the genetic basis for toxicity in the opportunistic pathogen *Staphylococcus aureus*, which used a genome-wide association study (GWAS) to identify potential genetic loci which may be associated with the toxic phenotype. This work revealed several novel genes which when disrupted caused a significant loss in toxicity in *S. aureus*, helping to further our understanding of the highly complex genetic regulation of virulence factors in this bacterium (Laabei et al., 2014; Laabei et al., 2015).

1.2.1 Genome-wide association studies (GWAS)

A genome-wide association study is conducted using a large dataset of genomic data, traditionally consisting of single-nucleotide polymorphisms (SNPs) present throughout the whole genome of an organism, and a corresponding set of quantifiable phenotypic data. In its most basic form, GWAS is a simple statistical test for association between particular loci in the genome and the phenotype of interest. It is important to note that GWAS can only identify statistical associations between genotype and phenotype, and causal relationships have to be verified through further investigation. It is well known that GWAS produces a high number of false associations with the phenotype (false positives) and can also wrongly exclude significant associations (false negatives), therefore functional verification is a crucial aspect of this research if meaningful results are to be obtained. Nonetheless, using this top-down approach can be a useful tool for identification of what would otherwise remain unknown effectors of a phenotype.

GWAS was originally designed to identify genetic polymorphisms associated with human disease, with the first published study looking to identify single nucleotide polymorphisms (SNPs) associated with myocardial infarction (Ozaki et al., 2002). This study identified a single SNP which was statistically associated with this condition and further experimental testing confirmed the finding. Another well-known example of GWAS in studies of human disease was the discovery of a common breast cancer risk allele among carriers of the *BRCA1* gene (Antoniou

et al., 2010; Kraft & Haiman, 2010). As a result of this research, screening tests now look for the BRCA1 mutation in patients which is associated with a higher risk of developing breast cancer. The success of these studies demonstrates the potential impact this GWAS method can have on medical research, and so more recent research has looked to apply this approach to the study of bacterial phenotypes relevant to disease such as host specificity (Sheppard et al., 2013), toxicity (Laabei et al., 2014) and antibiotic resistance (Farhat et al., 2013; Alam et al., 2014).

1.2.1.1 Units of genetic variation used in GWAS

GWAS has traditionally used single nucleotide polymorphisms (SNPs) as the unit of genetic variation to investigate a phenotype (Bush & Moore, 2012). These are single base changes that can be either synonymous (do not cause a change in the resulting amino acid) or non-synonymous (cause a change in the resulting amino acid). As described above this has yielded some important discoveries in genetic contributors to human disease. With genome sequencing data becoming more readily available and bioinformatics techniques continue to advance, researchers are beginning to use a variety of different measures of genetic variation, especially in bacterial GWAS, in an effort to increase the accuracy of GWAS predictions. For example, Sheppard et al. (2013) developed a method of dividing the genome into short segments of around 30bp in length for use as the unit of genetic variation; Laabei et al (2014) used both SNPs and indels in their GWAS method; and a number of studies have instead used presence or absence of whole genes (van Hemert et al., 2010; Chaston et al., 2014; Salipante et al., 2015).

1.2.1.2 Linkage disequilibrium

One of the major criticisms of GWAS is the high number of false positive and false negative associations this method identifies. The identification of false positive results is in part due to linkage disequilibrium, referring to the degree to which SNPs are inherited together during recombination in non-random proportions (Slatkin, 2008; Bush & Moore, 2012; Sved & Hill, 2018). Highly clonal populations of bacteria, in comparison to human populations, have high levels of genome-wide linkage disequilibrium because recombination does not take place as frequently (Read & Massey, 2014). For example, *Mycobacterium tuberculosis* is a highly clonal species, and in bacterial GWAS this increases the difficulty of identifying a causative SNP among a high number of statistically significant

associations (Chen & Shapiro, 2015). Bacteria with high levels of recombination are much more desirable for GWAS, because this effect of linkage disequilibrium is reduced, although not eliminated. *Streptococcus pneumoniae* is one such species of highly recombining bacteria, with shorter, more localised regions of linkage disequilibrium across the genome (Chen & Shapiro, 2015).

1.2.1.3 Direct and indirect associations

Due to this effect of linkage disequilibrium, it is essential to distinguish direct from indirect associations in GWAS results. Direct associations refer to the specific causative loci which directly affect the phenotype. Indirect associations are loci which are identified as being statistically significantly associated with the phenotype but are in fact in linkage disequilibrium with the causative SNP and therefore not an effector of the phenotype. Distinguishing these two types of associations requires functional verification of the effect a significantly associated loci has on the phenotype of interest. In bacteria, this can be done by genetic manipulation of associated loci and subsequent assaying of the phenotype. Though potentially labour and time intensive, with the introduction of mutant libraries such as the Nebraska Transposon Mutant Library for *Staphylococcus aureus* (Fey et al., 2013), functional verification is becoming increasingly feasible in a high-throughput manner.

1.2.1.4 Population stratification

Population stratification is another influence on statistical associations made through GWAS which can be responsible for identification of false associations with the phenotype – both false positives and false negatives. The problem of population stratification refers to the nature of bacterial populations to be strongly segregated into distinct genetic lineages that are highly influenced and maintained through the action of natural selection (Earle et al., 2016). Natural selection can be a strong driver of the propagation of antibiotic resistance and virulence phenotypes, for example, through a bacterial population (Chen & Shapiro, 2015). Such differences between genetically distinct populations can confound GWAS results if not considered in the analysis because they account for the observed variability in a phenotype (Earle et al., 2016).

For example, genetic variant A may be causative of phenotype B in one genetically distinct subset of a bacterial population; if the GWAS analysis is conducted on the population as a whole, it will not pick up this causative

association because the effect of variant A on the phenotype in this subset will be masked by the wider population in which it does not affect the phenotype (it wil be a false negative). However, if population structure is accounted for in the analysis it will become clear that in a specific subset of the population, variant A has a strong association with the phenotype.

On the other hand, if population stratification is too tightly controlled, genetic variants that are unrelated to the phenotype but only present in distinct subsets of the population that exhibit the phenotype may be falsely identified as causative loci (they will be false positive associations). As with many aspects of GWAS, controlling for population stratification relies on a balance between excluding loci that are likely to confound the analysis and including loci that are present only in a subset of the population but may be directly causative of a change in the phenotype.

Work by Maury et al. (2016) is a good example of the increased power of association studies which appropriately controls for population stratification. In this work, focussing on population structure in an association study of the foodborne pathogen *Listeria monocytogenes* led to identification of virulence attributes which were specifically relevant to either central nervous system listeriosis or maternal-neonatal listeriosis, both of which are clinically important. Listeriosis is a serious and sometimes deadly disease, so understanding the virulence factors which specifically contribute to different types of disease caused by clinical *L. monocytogenes* isolates could prove to be extremely useful in the development of better treatments. Here, accounting for population structure was shown to be important in increasing the power of GWAS analysis.

1.2.1.5 Corrections for multiple testing

A further aspect of GWAS to consider is the effect of multiple testing. The threshold of significance for an individual statistical test is usually set at p=0.05, meaning that the probability of an association being incorrectly observed if the null hypothesis is true is 5%. The null hypothesis here would be that there is no association between an individual genetic locus and the observed phenotype, and results where this hypothesis is rejected when it is true are referred to as type I errors, or false positive results. In a small dataset, for example of 20 results, this would mean only 1 result would be false positive. However, as the dataset increases in size, the number of false positive results increases simultaneously, so in a dataset of 100, 5 results will be false positive, in a dataset

of 1000, 50 will be false positive, and so on. GWAS can involve potentially millions of statistical tests for association in a single dataset, resulting in a huge number of results for which the null hypothesis is incorrectly rejected under a significance threshold of p=0.05. One way to overcome this effect is to apply corrections for multiple testing, examples of which include the Bonferroni correction, the false discovery rate and permutation of the dataset.

The Bonferroni correction is a well-known method of correcting for multiple tests; it is a simple calculation designated ‘α/n’ where ‘α’ refers to the threshold at which a p-value is considered significant (eg. 0.05) and ‘n’ refers to the number of tests conducted. This adjusts the significance level for each individual test to a much more stringent threshold but maintains the study-wide error rate at 0.05, or 5% (Perneger, 1998). However, this type of correction is not always valid in biological science; for example, the Bonferroni correction assumes that each test for association is independent of the others, but due to linkage disequilibrium, particularly in highly clonal populations, this is not always the case (Bush & Moore, 2012). In some cases, this type of correction can be too stringent and result in an increase in type II errors, also known as false negative results, whereby the null hypothesis is accepted when it should be rejected (Perneger, 1998; Armstrong, 2014).

An alternative to the Bonferroni correction is the false discovery rate, first proposed by Benjamini & Hochberg (1990; 1995). Instead of adjusting the significance threshold to reduce the likelihood of producing false positive results, the false discovery rate estimates the proportion of false positive results within the dataset, correcting for this to determine the number of true results among those which are statistically significant (Bush & Moore, 2012). Control of the false discovery rate has proved a useful tool in the analysis of GWAS data to adjust for the high number of statistical comparisons and in turn reduce the incidence of false positive results (van den Oord, 2007; Brzyski et al., 2017).

A third approach to correcting for multiple tests is permutation of the dataset. A permutation test involves scrambling the original dataset to remove the relationship between genotype and phenotype, creating a situation whereby the null hypothesis is true and there is no biological basis for the alignment between genotype and phenotype. Following permutation, the GWAS analysis is conducted again on the scrambled dataset (note this is not a secondary analysis), and this is repeated thousands of times so that any association found

between genotype and phenotype is purely due to chance. The p-value for each locus is then recalculated after all permutations have been completed to determine the likelihood of an association being made by chance alone, taking into account the average p-value under the null hypothesis (permutation) and the number of permutations conducted. This strengthens the findings of GWAS by providing empirically-derived p-values and demonstrating that significance is not an artefact of chance. Permutation of the dataset and subsequent calculation of empirically derived p-values can be done in numerous ways depending on the hypothesis being tested. For example, permutation could be applied to SNP data, to whole genes or to the phenotypic data, or to a combination of these (Guo et al., 2009; Cantor et al., 2010; Backes et al., 2014; Davenport et al., 2015; Brynildsrud et al., 2016). While permutation in principle is a straightforward approach to correcting for multiple tests, the number of permutations required to obtain statistically meaningful results can be computationally intensive, which is a potential drawback to this approach.

1.2.1.6 Filtering of genotypic data

Like permutation, GWAS analysis itself can be a computationally intensive process. One method to reduce the number of comparisons which are made in the first place is to filter genotypic data to eliminate loci with high frequencies of missing data and low frequencies of genetic polymorphisms. As has been discussed with many other aspects of GWAS, filtering of the dataset could lead to exclusion of potentially important loci from the analysis, so it is important that reasonable thresholds are set for exclusion of data. A threshold referred to as the genotype frequency can be set to eliminate loci with, for example 90% missing data, and a threshold referred to as the minor allele frequency can be set to eliminate loci with, for example, less than 5% polymorphisms. These thresholds can be manipulated to the needs of the researcher and can add power to the analysis by reducing the overall number of comparisons made.

Other approaches could include removing loci which cause only synonymous mutations, or grouping all mutations within a genetic feature, such as a gene or intergenic region, and treating each group as one statistical comparison (Read & Massey, 2014). However, eliminating loci from the genotypic dataset could potentially lead to the elimination of biologically important loci from the analysis, resulting in an increase in false negative results. Highly conserved loci, for example, will have very few SNP changes, and may be eliminated from the

analysis by the minor allele frequency threshold, as would rare mutations that may have a strong effect on the phenotype. This step must therefore be a trade-off between increasing statistical power of the analysis, and potentially losing biologically meaningful results.

1.2.1.7 Visualisation of GWAS results

Despite statistical correction for multiple testing and prior filtering of genomic data, false positive and negative results cannot be completely eliminated from GWAS analysis. To some degree, it is therefore important to judge GWAS results in terms of their biological relevance as well as statistical significance. One way to do this is to graph the results to visualise where the peaks and clusters of significant associations are along the genome; this is usually done using the Manhattan scatterplot format (see figure 1.1), where the -log10 of each p-value is plotted and statistically significant loci appear as peaks along the X-axis, comparable to the Manhattan skyline from which the graph gets its name (Zhang et al., 2014). Visualising the data in this way does not allow discrimination between direct and indirect associations, but it does allow identification of clusters of significant associations which are likely to contain a causative SNP.

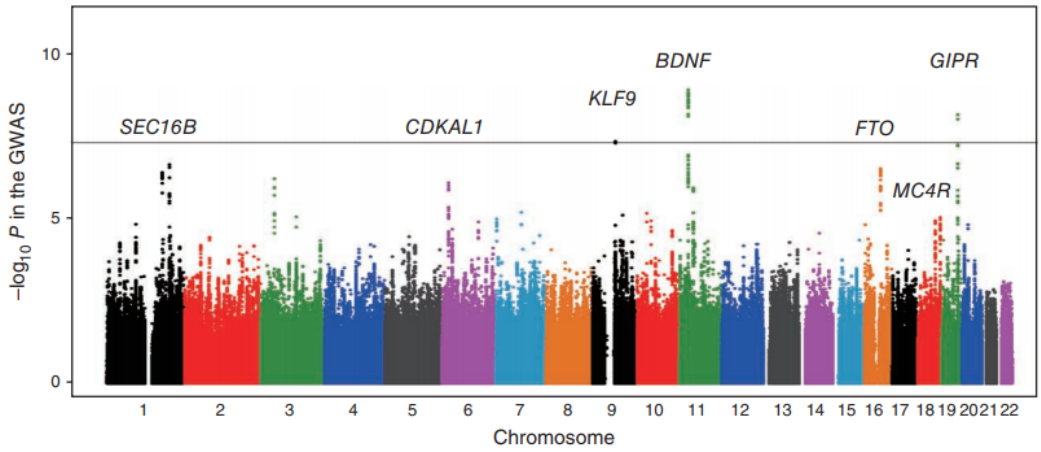


Figure 1.1. Example of a Manhattan scatterplot. Position of the locus on the chromosome is plotted on the X axis, against the -log^{10} of the p-value, indicating significance on the Y axis. The horizontal grey line represents the threshold for statistical significance. Clusters of statistically significant loci associated with the phenotype, in this case body mass index in east Asian populations, are seen as peaks on the plot. From Okada et al. (2012), reused with permission from Springer Nature.

Recent work has led to the development of interactive applications such as Phandango (Hadfield et al., 2017) for the visualisation of GWAS data beyond the

static Manhattan plot. Phandango allows researchers to visualise the output from multiple different genomic analysis methods and explore regions of significance in a more intuitive way, combining GWAS results with the reference sequence, phylogenetic data and genome annotations, all of which can be tailored to the needs of the user. From this, inferences can be made regarding the likely biological function of a genetic locus of interest in relation to the phenotype, allowing prioritisation of significant GWAS results based on both biological relevance and statistical significance.

1.2.1.8 Investigating bacterial phenotypes using GWAS

GWAS is increasingly being applied to studies of bacterial phenotypes, particularly in relation to antibiotic resistance and pathogenesis. The first published bacterial GWAS was conducted by Sheppard et al. in 2013 and investigated genetic factors in *Campylobacter jejuni* that are associated with adaptation to a specific host. The authors found a region of seven genes which were specifically associated with adaptation to cattle and frequently absent in strains which were isolated from chickens and wild birds. Three of these genes were found to be involved in vitamin B_5 biosynthesis, and the authors hypothesise that they may contribute to specific adaptation to the cattle diet. Host adaptation is an important aspect of bacterial pathogenicity, as transmission dynamics between hosts can influence the survival and persistence of a pathogen (Dobson, 2004), and this research presents GWAS as a useful tool for investigation of bacterial pathogenicity using a genome-wide approach.

A number of studies since then have focussed on identifying specific genetic loci in bacteria which contribute to antibiotic resistance (Farhat et al., 2013; Alam et al., 2014; Chewapreecha et al., 2014). For example, Chewapreecha et al. (2014) used GWAS to pinpoint specific polymorphisms within pneumococcal 'mosaic genes' which contribute to resistance to β-lactam antibiotics. Mosaic genes are regions of DNA containing polymorphisms from the cell's original DNA in combination with those from donor DNA obtained through recombination (Boc & Makarenkov, 2011). While previous studies had identified these blocks of recombination as associated with β-lactam resistance, Chewapreecha et al. used GWAS to refine smaller loci of discrete or linked SNPs which contribute to the phenotype.

The GWAS approach has also been expanded from identifying effector polymorphisms in isolation to identifying epistatically interacting loci; these being

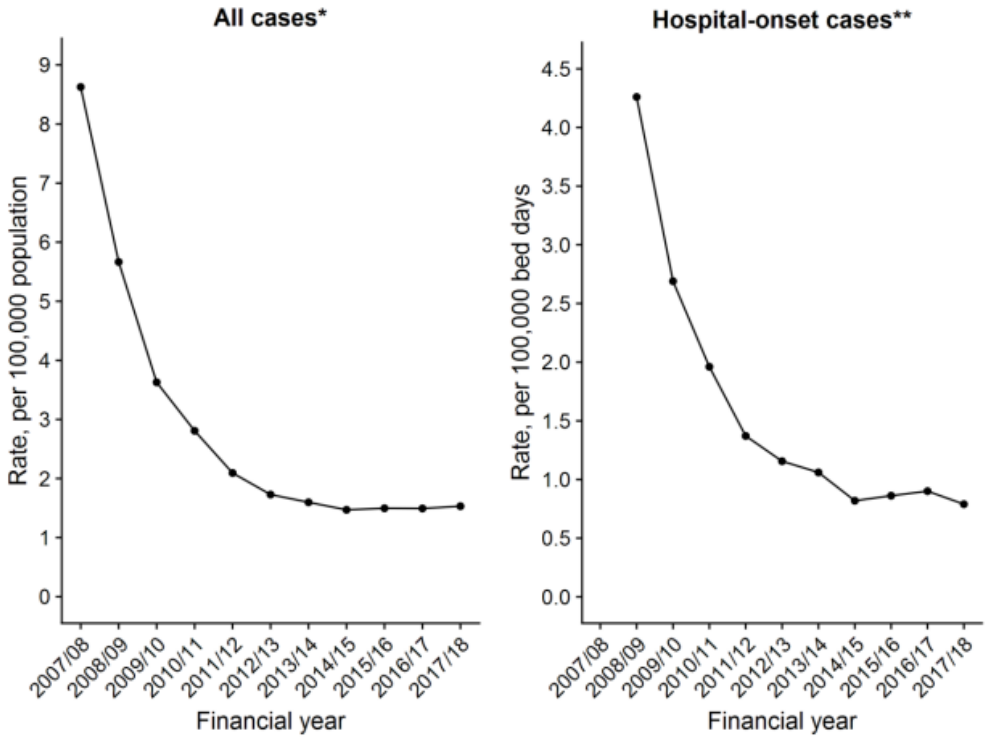
loci that alone are not significantly associated with a change in the phenotype but do become significantly associated when interacting with other loci, whereby one gene masks or modifies the effects of another (Laabei et al., 2014; Yokoyama et al., 2018; Schubert et al., 2019). For example, work by Laabei et al. (2014) identified SNPs in several genes including *ileS* and *mreC* which were found to be epistatically interacting with other loci to affect toxicity in *S. aureus* with high statistical significance. Follow-up work by Yokoyama et al. (2018) to functionally verify the role of the *ileS* gene in relation to toxicity demonstrated that mupirocin resistance, which is conferred by a mutation in the *ileS* gene, affects toxin production when in combination with specific polymorphisms at other loci. In turn this was found to affect the fitness of *S. aureus* to the extent that this mutation has not spread through the population as much as would be expected for one which confers antibiotic resistance.

All the examples given above demonstrate the broadening application of GWAS in expanding our understanding of the genetic factors which contribute to bacterial pathogenesis. Traditionally this line of investigation has taken a bottom-up approach, whereby the function of individual genes is disrupted and the resulting phenotype is studied to determine the role of the gene. With the increasing availability of genome sequences and improvements in bioinformatic methods, it is becoming increasingly possible to take a top-down approach to this research. Despite its flaws in terms of the generation of false positive and false negative results, and the extensive amount of quality control which must be performed on the dataset, GWAS has been shown to be a useful tool in this top-down approach, allowing researchers to start with a large dataset of whole genome sequences, and narrow down regions of interest for targeted investigation of specific loci in relation to a phenotype of interest.

1.3.0 Introduction to *Staphylococcus aureus*

Strains of methicillin-resistant *Staphylococcus aureus* (MRSA) are perhaps best-known among the *S. aureus* species, having been a persistent cause of hospital- and community-acquired infections in the UK over recent decades and has been identified as a public health priority within the European Union (Köck et al., 2010). Gould (2006) describes MRSA as “by far the most significant antibiotic-resistant pathogen we have ever encountered” and following a dramatic rise in MRSA bacteraemia over the turn of the century surveillance of MRSA bacteraemia was made mandatory (Johnson et al., 2012). A concerted effort has since been made

to improve hygiene practices in hospitals for the prevention of MRSA and as a result the prevalence of MRSA bacteraemia has substantially declined over recent years (see figure 1.2).

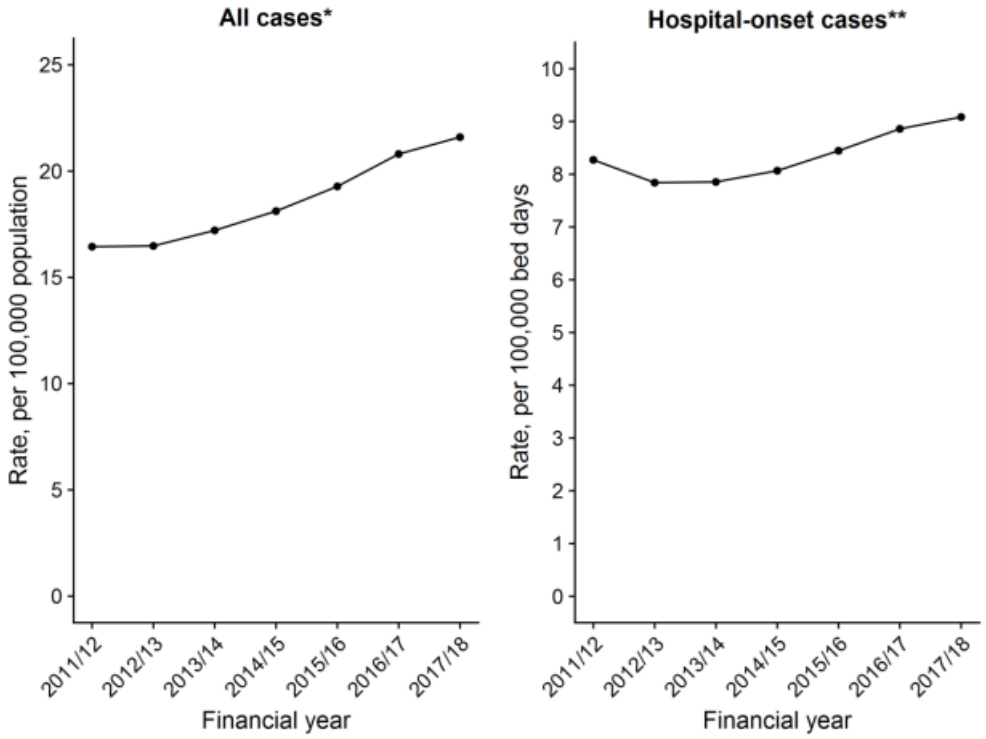


* * Mid-year population estimates for 2017/18 were not available at time of publication and so population data for 2016/17 were used as a proxy.

** Bed day data were not available for quarter 4 of FY 20152017/16 18 (January to March, 20162018). As a result, the 20162017/17 18 bed day data is an aggregate of quarters 1, 2 and 3 of 2016 2017/18 and quarter 4 of 2016/17.

Figure 1.2. Trends in the rate of MRSA bacteraemia in England. Illustration of the reduction in prevalence of MRSA bacteraemia from 2007-2018. Figure from Public Health England (2018a), reused under the Open Government Licence v3.0.

While MRSA bacteraemia is shown to be in decline, cases of methicillin-susceptible *Staphylococcus aureus* (MSSA) bacteraemia have been on the rise since surveillance was made mandatory in 2011 (see figure 1.3).



*Mid-year population estimates for 2017/18 were not available at time of publication and so population data for 2016/17 were used as a proxy.

** bed day data were not available for quarter 4 of FY 2017/18 (January to March, 20172018). As a result, the 2017/18 bed day data is an aggregate of quarters 1, 2 and 3 of 2017/18 and quarter 4 of 2016/17.

Figure 1.3. Trends in the rate of MSSA bacteraemia in England. Illustration of the increase in prevalence of MSSA bacteraemia from 2011-2018. Figure from Public Health England (2018a), reused under the Open Government Licence v3.0.

It is interesting from an evolutionary and epidemiological perspective that MSSA is on the rise while MRSA is in decline. It seems there is a reversal of roles between methicillin-resistant and -sensitive strains, with MSSA shown to replace MRSA as the predominant cause of nosocomial infections. Furthermore, in America for example, MRSA has become a prevalent cause of *S. aureus* infection in the community (David et al., 2011). Community-acquired disease caused by *S. aureus* was first reported in injecting drug users in 1981 (Saravolatz et al., 1982) and this was later followed by the deaths of four young children from MRSA infection, despite having no apparent risk factors for disease (Centers for Disease Control and Prevention, 1999). Since these early reports, the incidence of community-acquired *S. aureus* infections has risen worldwide, with methicillin-resistant isolates emerging as the predominant cause of staphylococcal disease in the community (Voyich et al., 2006; Boucher & Corey, 2008; Gordon & Lowy, 2008).

It was noted by David et al. (2011) that MSSA isolates causing clinical infections in hospitals rarely carried genes for the Panton-Valentine leukocidin (PVL) toxin, which contributes to increased virulence in isolates that carry it (although it is not the singular determinant of increased virulence [Voyich et al., 2006]). In contrast, most of the community-acquired MRSA (CA-MRSA) isolates identified by David et al. were PVL-positive. The authors imply based on these findings that nosocomial MSSA strains are less virulent than the PVL-positive CA-MRSA isolates, which have to invade an otherwise healthy host with a fully functional immune system in most cases. MSSA isolates, on the other hand, have access in the hospital environment to immune-compromised patients who are less competent to resist the onset of staphylococcal disease and therefore do not need such high levels of virulence in order to colonise and invade a vulnerable host. This, in combination with the intense effort to reduce the prevalence of MRSA in hospitals over recent decades, could explain the increase in cases of MSSA infections.

Interestingly, however, high levels of toxicity do not necessarily correlate to more severe disease. This principle was demonstrated in work by Laabei et al. (2015), in which it was shown that there is an inverse correlation between bacterial toxicity and disease severity. These results are consistent with the work of David et al. (2011), which demonstrated that the supposedly more virulent CA-MRSA strains were predominantly responsible for less severe skin and soft tissue infections, while patients infected with MSSA were more likely to develop invasive disease, although this data represents a static time point and disease outcomes are not reported.

Laabei et al. explain this observed trade-off between bacterial toxicity and disease severity by demonstrating that low-toxic *S. aureus* isolates have higher relative fitness in human serum than high-toxic isolates, and showed that high-toxic isolates have a competitive advantage at the population level and are more likely to transmit successfully between hosts. Isolates of lower toxicity, on the other hand, are more consistently found in bacteraemia patients, which is an evolutionary dead end for *S. aureus* in the event the disease is fatal to the host. Nonetheless, follow-up work by Recker et al. (2017) demonstrated that once in the bloodstream a low-toxic isolate is less likely to lead to mortality than a high-toxic isolate, as would be expected.

MRSA remains a huge concern for clinicians in this era of antibiotic resistance, and other drug-resistant variants such as glycopeptide-intermediate *S. aureus* (GISA) and vancomycin-resistant *S. aureus* (VRSA) are also a particular threat should they become more prevalent causes of staphylococcal disease (Gould, 2006). However, the rise in cases of MSSA infection should not be ignored, and epidemiological monitoring of the fluctuations in disease prevalence of both methicillin-resistant and -sensitive *S. aureus* will be important in designing future prevention strategies against staphylococcal disease.

1.3.2 *S. aureus* virulence factors

In causing a range of diseases from milder skin and soft tissue infections, to the more invasive diseases such as pneumonia, meningitis and septicaemia, *Staphylococcus aureus* employs an army of virulence factors (see figure 1.4 below for examples). Bacterial virulence is highly dependent on genetic and environmental determinants for both host and pathogen, but *S. aureus* nonetheless is well equipped to cause disease in the human host due to this extensive array of virulence factors.

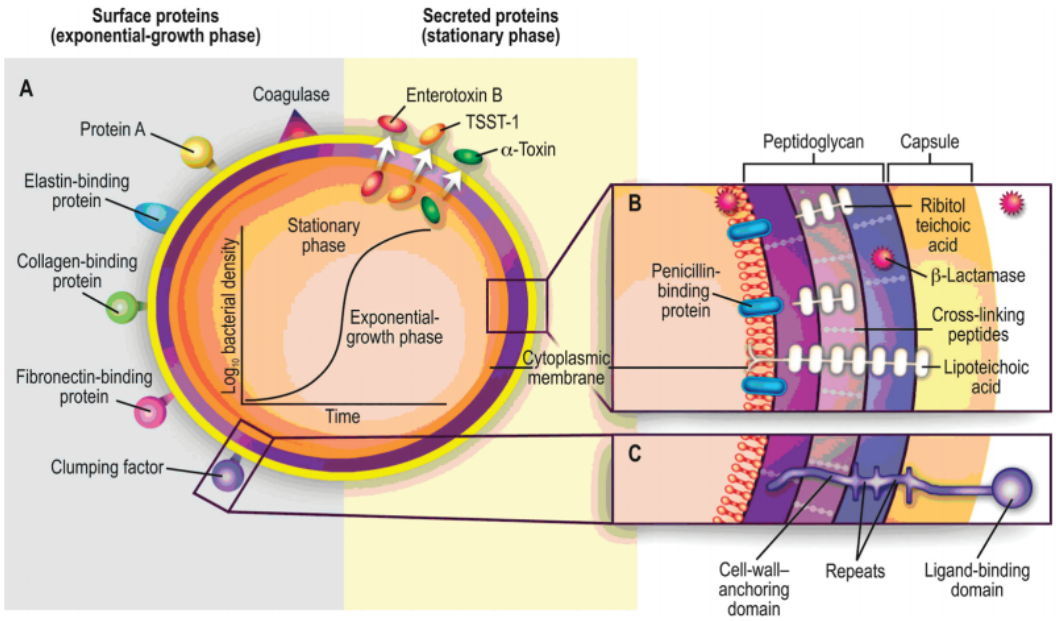


Figure 1.4. Virulence factors of *Staphylococcus aureus*. (A) Surface and secreted proteins. (B) and (C) Cross-sections of the cell envelope. TSST-1 – toxic shock syndrome toxin 1. Figure reproduced with permission from Lowy (1998), Copyright Massachusetts Medical Society.

1.3.2.1 Adhesins

S. aureus produces a wide array of adhesive proteins that contribute to its ability to colonise host cells and tissues in the early stages of infection (Gordon & Lowy, 2008). A number of these adhesins belong to a large group of surface proteins that are collectively referred to as 'microbial surface components recognizing adhesive matrix molecules' or MSCRAMMs. MSCRAMMs are a key part of the colonization stage of infection and while most are adhesive proteins, some also play a role in evasion of host immunity (Foster et al., 2014). MSCRAMMs are more specifically defined by Foster et al. as a family of proteins with a common mechanism of ligand binding mediated by two adjacent subdomains containing IgG-like folds. Examples of adhesins within this group include clumping factors A and B (ClfA/ClfB), fibronectin-binding proteins A and B (FnBPA and FnBPB) and the collagen adhesin Cna.

ClfA specifically adheres to fibrinogen in the extracellular matrix of host tissues to facilitate colonisation (O'Connell et al., 1997), and promotes clumping of *S. aureus* cells in the bloodstream by binding fibrinogen within the plasma (Ganesh et al., 2008). ClfB is similar in structure to ClfA but binds a different region of host fibrinogen; Eidhin et al. (1998) propose that the two clumping factor proteins act together to enhance binding to fibrinogen within the blood stream. A role has also been proposed for ClfB in adherence to the squamous epithelial cells of host nasal passages to promote specific colonisation of this niche (Mulcahy et al., 2012) and it has been demonstrated that this can be mediated by the structural protein cytokeratin 10 (Walsh et al., 2004; Ganesh et al., 2011) and the cell envelope protein loricrin of squamous epithelial cells (Mulcahy et al., 2012). Clumping factor mutants have been shown to be attenuated in adherence to host cell components and tissues (Eidhin et al., 1998; Mulcahy et al., 2012), and further roles have also been demonstrated for clumping factors in evading phagocytosis (Higgins et al., 2006), in binding to platelets to initiate infective endocarditis (Siboo et al., 2001) and in binding to biomaterials used in surgical implants (Vaudaux et al., 1995).

The fibronectin-binding proteins FnBPA and FnBPB, as their name suggests, bind fibronectin within the blood plasma and the extracellular matrix of host tissues, and also contribute to adherence to surgical implants (Foster & Höök, 1998; Foster et al., 2014). Work by Peacock et al. (1999) demonstrated that the FnBPs are the dominant adhesins involved in binding of *S. aureus* to human

endothelial cells and are also key to internalisation of *S. aureus* into endothelial cells to promote invasion during disease. Massey et al. (2001) extended this work to elucidate the fibronectin-binding regions of FnBPA, which are described as forming a bridge between bacterial cells and endothelial cells that is connected via fibronectin. This work also demonstrated that FnBPA is the most important factor responsible for internalisation of *S. aureus* into endothelial cells.

Collagen adhesin Cna adheres specifically to collagen-rich tissue (Foster et al., 2014) and has been shown to contribute to virulence in models of bone infection, endocarditis and septic arthritis (Patti et al., 1994; Hienz et al., 1996; Elasri et al., 2002). For example, Patti et al. (1994) demonstrated that mutants lacking the *cna* gene were unable to bind collagen or express a collagen adhesin and were reduced in ability to adhere to cartilage. In a subsequent animal model of septic arthritis only 27% of mice treated with *cna*-deficient mutants developed symptoms of disease compared to 70% of mice treated with *cna*-positive strains.

Adhesins such as those described above have been identified as potential candidates for development of a staphylococcal vaccine, and work has shown that immunisation in mouse models with MSCRAMM proteins including ClfB and fibrinogen binding proteins can protect against initial colonisation, thereby protecting against onset of disease (Mamo et al., 1994; Schaffer et al., 2006). Recombinant ClfA and Cna fragments have also been shown to reduce the severity of septic arthritis in mouse models, with mortality significantly reduced in vaccinated hosts (Nilsson et al., 1998; Josefsson et al., 2001).

A further critical role of adhesins in virulence is their contribution to formation of biofilms in the initial stages of infection (Rohde et al., 2007; O'Neill et al., 2008; Heilmann, 2011), a process usually associated with indwelling medical devices such as catheters or surgical implants (Otto, 2008). Biofilms have been shown to aid resistance to antibiotic action by both providing a physical barrier to cells in the centre of the biofilm and because cells in a biofilm typically grow slowly with reduced metabolism which facilitates tolerance of antibiotic treatment (Marrie et al., 1982; Archer et al., 2011). These principles also apply to resistance against action of the host immune system (Costerton et al., 1999). Due to the persistence of biofilms within the host once established, they can become a reservoir for dissemination of *S. aureus* to other sites within the body (Otto, 2008).

1.3.2.2 Immune evasion

Once staphylococci have colonised the host, they must evade the action of the host immune system if they are to persist and spread. Infection with staphylococci stimulates a strong inflammatory response, so *S. aureus* must be prepared to actively evade host immunity, which it does through the deployment of several virulence factors dedicated to immune evasion. Examples of some of these immune evasion proteins are shown in figure 1.5 below.

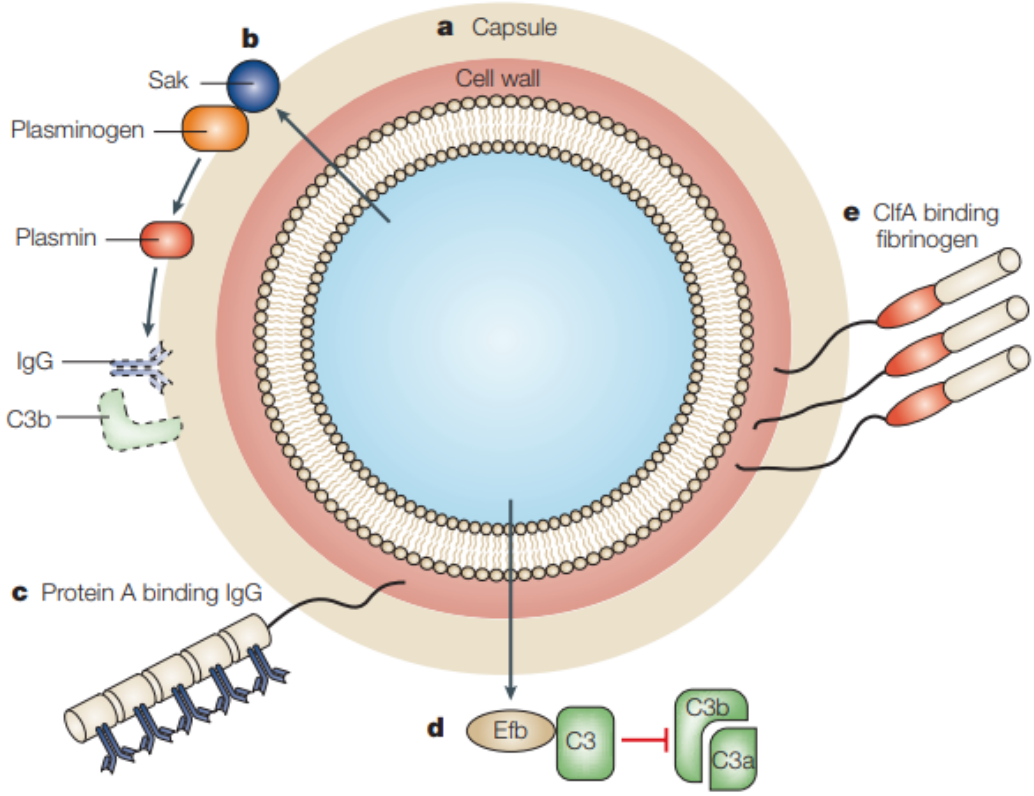


Figure 1.5. Mechanisms by which *S. aureus* avoids opsonophagocytosis. (A) Capsular polysaccharide compromises neutrophil access to bound complement and antibody. (B) Extracellular staphylokinase (Sak) activates cell-bound plasminogen and cleaves IgG and C3b. (C) Protein A binds the Fc region of IgG. (D) Fibrinogen-binding protein Efb binds complement factor C3 and blocks its deposition on the bacterial cell surface, thereby preventing complement activation beyond C3b attachment and inhibiting opsonisation. (E) Clumping factor A (ClfA) binds the γ-chain of fibrinogen. Figure taken from Foster (2005), reproduced with permission from Springer Nature.

The polysaccharide capsule provides a physical barrier to components of host immunity. For example, it prevents access of neutrophils to serum opsonins such as C3b which bind to the cell surface beneath the polysaccharide layer (Verbrugh et al., 1982), reducing uptake of staphylococcal cells by neutrophils and thereby interfering with opsonophagocytosis. In doing so, it has been shown that this

increases staphylococcal virulence, and in turn disease severity, in mouse models of infection (Nilsson et al., 1997; Thakker et al., 1998). Most clinical *S. aureus* isolates are found to express one of either serotypes 5, 8 or 336 capsular polysaccharide, which forms a thin 'microcapsular' layer around the cell (Foster, 2005). Out of a total of 11 serotypes identified thus far, serotypes 5 and 8 have been specifically associated with increased virulence in mouse models of infection (Nilsson, 1997; Luong & Lee, 2002). For example, work by Nilsson et al. (1997) showed that strains positive for capsular serotype 5 (CP5+) caused a higher frequency and severity of arthritis in a murine model, while *in vitro* assays suggested that CP5+ strains were phagocytosed less by macrophages and once phagocytosed were killed less efficiently than their isogenic capsule-deficient mutants. A number of studies have also shown that *S. aureus* strains of capsular serotypes 5 and 8 are able to resist phagocytosis by human polymorphonuclear leukocytes (Peterson et al., 1978; Verbrugh et al., 1982; Karakawa et al., 1988).

Along with the polysaccharide capsule, another well-characterised contributor to immune evasion is staphylococcal protein A (SpA), a cell wall-anchored protein which binds to the Fc region of host immunoglobulin IgG. Binding of the Fc region renders the antibody unrecognisable to host neutrophils because it is in the incorrect orientation. As the bacterial cell becomes covered in these incorrectly oriented IgG molecules it is concealed from neutrophils, and in this manner evades opsonophagocytosis (Foster, 2005). Furthermore, this protein A-IgG complex then goes on to bind components of the complement system which are usually involved in opsonisation, as the variable region of IgG to which complement would normally bind remains exposed. Mutants deficient in protein A are phagocytosed much more efficiently than protein A-rich strains and exhibit reduced virulence in animal models of infection (Patel et al., 1987; Gemmell et al., 1991; Palmqvist et al., 2002). The importance of IgG in the action of protein A was demonstrated by Peterson et al. (1977), who showed that protein A-rich strains were phagocytosed quicker in IgG-deficient serum compared to isogenic protein A-deficient strains.

Important roles have also been demonstrated for protein A in binding to tumour necrosis factor receptor 1 (TNFR1) on host epithelial cells, and in causing depletion of B lymphocytes. Binding of protein A to TNFR1 induces a strong inflammatory response in the respiratory tract and a signalling cascade that leads to recruitment of high numbers of neutrophils (Gómez et al., 2004). The rapid increase in inflammation causes increased tissue damage which is detrimental to

the host and can ultimately be responsible for the induction of staphylococcal pneumonia. Interference of protein A with B cells via binding to B cell receptors (BCRs) leads to rapid down-regulation of BCRs culminating in apoptosis (Goodyear and Silverman, 2003). Protein A can therefore promote the pathogenesis of *S. aureus* by evading the action of host immunity, directly interacting with host immune cells to inhibit their function and over-stimulating the immune response, depending on the stage of infection.

Several other virulence factors produced by *S. aureus* also interfere with phagocytosis by interacting directly with components of the human complement system. Examples include *Staphylococcus* complement inhibitor (SCIN), the extracellular fibrinogen-binding protein Efb, and the plasminogen activator protein staphylokinase (Sak). SCIN interacts directly with the C3 convertases C4b2a and C3bBb to stabilise the C3 convertase complex, inhibiting C3b deposition on the cell surface and thus preventing complement activation. This early intervention with complement ultimately prevents the formation of membrane attack complexes that would otherwise lead to lysis of the bacterial cell (Rooijakkers et al., 2005). Efb has been shown to bind the C3d fragment of C3b, inhibiting opsonophagocytosis, while simultaneously binding fibrinogen to interfere with platelet aggregation and wound healing (Palma et al., 1998; Lee et al., 2004). Staphylokinase binds host plasminogen that attaches to the bacterial cell surface, causing the formation of active plasmin that then goes on to cleave surface bound C3b and IgG, subsequently interfering with opsonophagocytosis (Foster, 2005; Bokarewa et al., 2006). It is also found to interact with the bactericidal peptide α-defensin produced by neutrophils to resist neutrophil killing (Bokarewa et al., 2006).

Other immune evasion proteins can act to indirectly protect the bacteria from the action of host immune cells. Coagulase, for example, binds prothrombin to clot the blood and mask the bacteria from the action of macrophages, indirectly protecting the pathogen from phagocytosis (McAdow et al., 2012). Interestingly, the golden pigment which is characteristic of *S. aureus* has also been identified as a virulence factor that contributes to immune evasion by protecting the bacteria from oxidative stress (Liu et al., 2005; Clauditz et al., 2006). Clauditz et al. demonstrated this through deletion of the *crtM* gene encoding the carotenoid pigment staphyloxanthin, which rendered the bacteria susceptible to oxidative stress and neutrophil killing.

1.3.2.3 Toxins

The third group of virulence factors produced by *S. aureus* are the toxins; proteins secreted by the bacterial cell that actively damage the host. Generally, *S. aureus* toxins can be divided into three distinct groups:

1) Membrane-damaging toxins
2) Toxins that interfere with receptor function only
3) Secreted enzymes that degrade host molecules and affect host defences

The army of staphylococcal toxins is extensive; the membrane-damaging toxins include α-toxin, Panton-Valentine leukocidin (PVL), the leukocidins LukDE and LukAB, γ-toxin and phenol-soluble modulins (PSMs). Toxins that interfere with receptor function include the enterotoxins, staphylococcal enterotoxin B and C (SEB/SEC), the superantigen toxic shock syndrome toxin-1 (TSST-1) and finally, the secreted, degradative enzymes include aureolysin, glutamyl endopeptidase SspA, β-toxin and lipase and nuclease enzymes. There is some crossover in the definition of a toxin here, with immune evasion proteins such as staphylokinase also being classified in the third group of secreted enzymes that degrade host molecules and defences.

α-toxin is one of the major cytolytic toxins of *S. aureus* and has been extensively studied over the years as a model for pore-forming toxins (Bhakdi & Tranum-Jensen, 1991; Dinges et al., 2000; Vandenesch et al., 2012). It exerts its toxic effect on host cells by binding as monomers to ADAM10 receptors on host cell membranes (Inoshima et al., 2011), which then oligomerise to form a heptameric complex and create a pore (Gouaux et al., 1997). Interestingly, binding of alpha-toxin to ADAM10 activates metalloprotease activity in epithelial cells, resulting in the cleavage of E-cadherin to which opportunistic *S. pneumoniae* can bind during infection (Devaux et al., 2019). The pore ultimately formed through the action of alpha-toxin allows indiscriminate passage of ions and small molecules up to a molecular weight of 4,000 daltons through the channel, causing osmotic swelling, and ultimately lysis of the target cell (Bhakdi & Tranum-Jensen, 1991). At sublytic concentrations, however, it was observed that α-toxin can act to influence a variety of host cells in the capacity of immune evasion (see figure 1.6 below), and it is becoming increasingly clear that the function of this and other toxins in relation to different stages of the bacterial life cycle is complex and incompletely understood (Berube & Wardenburg, 2013).

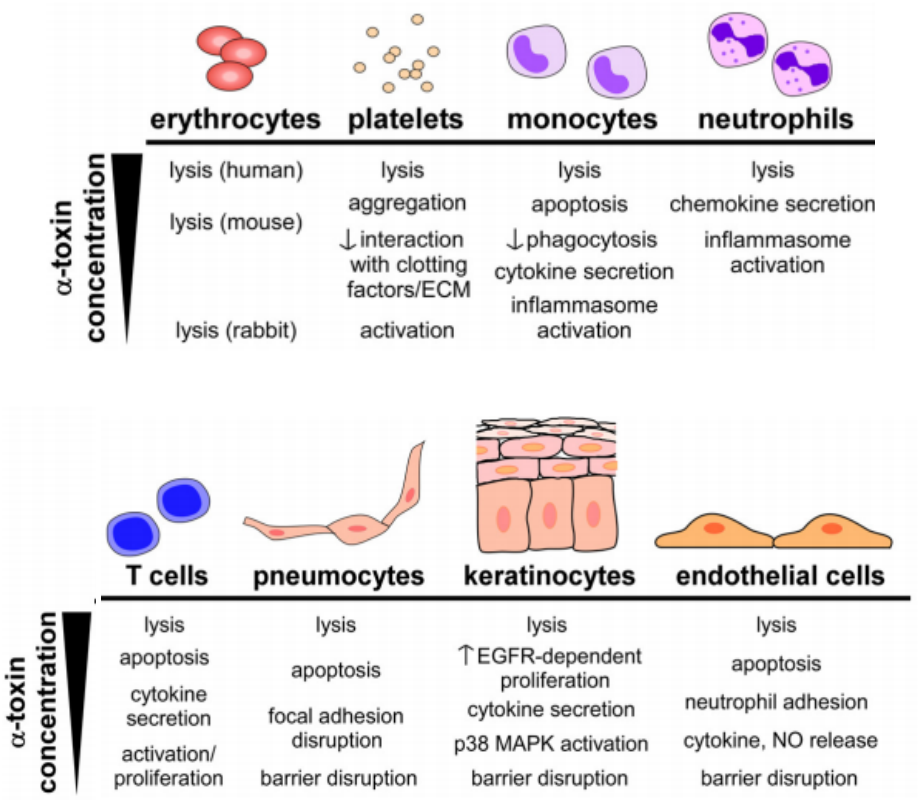


Figure 1.6. Cellular responses to intoxication by α-toxin. Multiple cell types are targeted by α-toxin, each displaying unique effects that are dependent on the relative concentration of toxin to which the cell is exposed. Figure taken from Berube & Wardenburg (2013) reused under the Creative Commons Licence 3.0.

The β-barrel family of pore-forming toxins to which α-toxin belongs also includes the Panton-Valentine leucocidin (PVL), the leukocidins LukDE and LukAB, and γ-toxin (Otto, 2014). PVL, like α-toxin, is also a receptor-mediated membrane damaging toxin which interacts with host C5aR and C5L2 receptors to form a pore in host cells (Spaan et al., 2013). PVL is of particular interest to researchers due to its apparent prevalence in community-acquired MRSA isolates, despite only being present in around 2-3% of *S. aureus* isolates overall (Dinges et al., 2000). The PVL locus is encoded on a bacteriophage and has been shown to be generally associated with skin infections, but in extreme cases can also be responsible for the development of necrotising pneumonia (Gillet et al., 2002). Whilst it has been suggested that PVL could be a stable genetic marker for CA-MRSA infection, and that it may have driven the CA-MRSA epidemic of clones such as USA300 (Vandenesch et al., 2003), more recent *in vitro* analyses of the contribution of PVL to *S. aureus* pathogenesis suggests only a minor role in pathogenesis (Badiou et al., 2010; Graves et al., 2010). Work by Voyich et al.

(2006) provides further convincing evidence that PVL is not a major contributor to virulence in CA-MRSA isolates; here the authors compared the virulence of PVL-positive and -negative CA-MRSA strains in mouse models of infection and found that strains deficient in PVL remained as virulent as PVL-positive strains. This finding was also supported by experiments using PVL knockout strains of USA300 and USA400, which caused comparable skin disease to the isogenic wild type strains. PVL may therefore be a potentially useful epidemiological marker for CA-MRSA strains in some cases but appears not to be the major determinant of virulence in these strains.

α-toxin and PVL are just two examples of well-characterised toxins of *S. aureus*. A variety of toxins are produced by this bacterium, as listed above, and whilst there are differences in the mechanisms by which they damage host cells, all have the potential to be potent effectors of virulence during staphylococcal disease.

Interestingly, work by Laabei et al. (2014) demonstrated that a greater amount of phenotypic variation is seen among the toxins of *S. aureus* when compared to adhesive proteins, which may account for the wide variety of diseases this bacterium can cause. Expanding on this, work by Laabei et al. in 2015 demonstrated that isolates extracted from skin and soft tissue infections are more often caused by high-toxic *S. aureus* strains, while isolates extracted from patients with life-threatening bacteraemia and septicaemia had comparatively low levels of toxicity, although among isolates causing bacteraemia, those with higher toxicity were more likely to cause mortality (Recker et al., 2017). Once a strain has gained access to the bloodstream it is not known whether it is capable of switching its level of toxicity, and strains are grouped into high- and low-toxic on the basis of *in vitro* assays. This inverse correlation between toxicity and disease severity was also observed by Nozohoor et al. in their work of 1998, in which lower levels of α-toxin were observed in strains causative of endocarditis compared to wound infection and nasopharyngeal carriage strains.

This difference between surface infections and invasive disease could be explained by the concept that the skin is a difficult physical barrier to break through and in which to maintain an infection, requiring higher virulence in the infecting isolate during the initial stages of colonisation and disease, whilst an isolate which has gained entry to the blood does not need to be highly virulent to cause potentially fatal disease. Importantly, this work by Laabei et al. (2014;

2015) suggests that staphylococcal toxicity may be tightly linked to disease severity, even if this link is an inverse relationship between the two. If we can better understand this phenotype, we may be better able both to predict the severity of disease and provide appropriate, personalised treatment. The work presented in chapter 3 of this thesis contributes to a better understanding of the genetic factors that contribute to toxicity in *S. aureus*. Furthermore, it is becoming increasingly clear that toxins are not just offensive tools used solely in the destruction of host cells, but that they contribute to various stages of the bacterial life cycle, including colonisation and asymptomatic carriage.

1.3.2.4 Changing perspectives on toxin function

Whilst much research has focused on the damaging effects of toxins in relation to their impact on host cells, Rudkin et al. (2017) emphasise that if we are to fully understand the role of toxins in the bacterial life cycle we must move away from a disease-centric perspective. While toxins are clearly advantageous to a bacterium in contributing to nutrient acquisition and transmission, from an evolutionary point of view, a highly toxic staphylococcal strain that causes fatal disease could find itself at an evolutionary dead-end upon the death of its host. So, what evolutionary advantage do such toxins provide?

Recent work has started to shift from a host perspective to the bacterial perspective, particularly focusing on the adaptable role of toxins in colonisation and asymptomatic carriage, alongside pathogenesis within a host. For example, it is now acknowledged that bloodstream staphylococcal isolates exhibit reduced toxicity compared to carriage and skin infection isolates (Nozohoor et al., 1998; Laabei et al., 2015), and that bacteraemia in a patient often stems from their own originally asymptomatic carriage isolate which transitions to cause invasive bloodstream infection (von Eiff et al., 2001). To understand why this happens, we must look further into the role of toxins in the initial stages of infection and into the evolutionary changes that occur within a population over a short space of time to effect changes in the infectious state.

Work by Das et al. (2016) has demonstrated that staphylococci undergo evolution within the human body, mutating the transcription factor repressor of surface proteins (*rsp*) to attenuate toxin production and transition from an asymptomatic, high-toxic carriage isolate to an attenuated-cytotoxicity, non-haemolytic phenotype capable of causing serious bloodstream infection. This ability of staphylococci to transition within the host based on the manipulation of toxin

expression demonstrates the important contribution of toxins at different stages of the bacterial life cycle, something we must work to understand if we are to gain a complete picture of the progression of staphylococcal pathogenesis.

It is important to note that host factors should not be forgotten in contributing to changes in the state of staphylococcal carriage isolates, as changes in immune status may be the initial trigger which allows changes in the bacterial phenotype to facilitate invasive infection. Nonetheless, further investigation of the genetic contributors to virulence, and specifically toxicity, in pathogens such as *S. aureus* will be useful in improving this understanding of the progression of staphylococcal pathogenesis during infection and disease, elucidating how toxicity is regulated throughout the bacterial life cycle to the detriment of the host.

1.3.3 Regulation of virulence determinants in *S. aureus*

The individual contribution of certain virulence regulators, such as the accessory gene regulator (*agr*) and the staphylococcal accessory regulator (*sarA*) have been extensively studied in *S. aureus* to characterise their role in virulence. However, the epistatic interactions between the broad range of genes which can influence virulence in *S. aureus* are less well characterised, and as yet it is not possible to produce a generalised model of the virulence regulatory network of this bacterium that is applicable to all strains. Furthermore, we do not yet have a complete picture of the network of genes which contribute individually to staphylococcal pathogenesis. Priest et al. (2012) demonstrate the complexity of the *S. aureus* virulence regulatory network in figure 1.7 below.

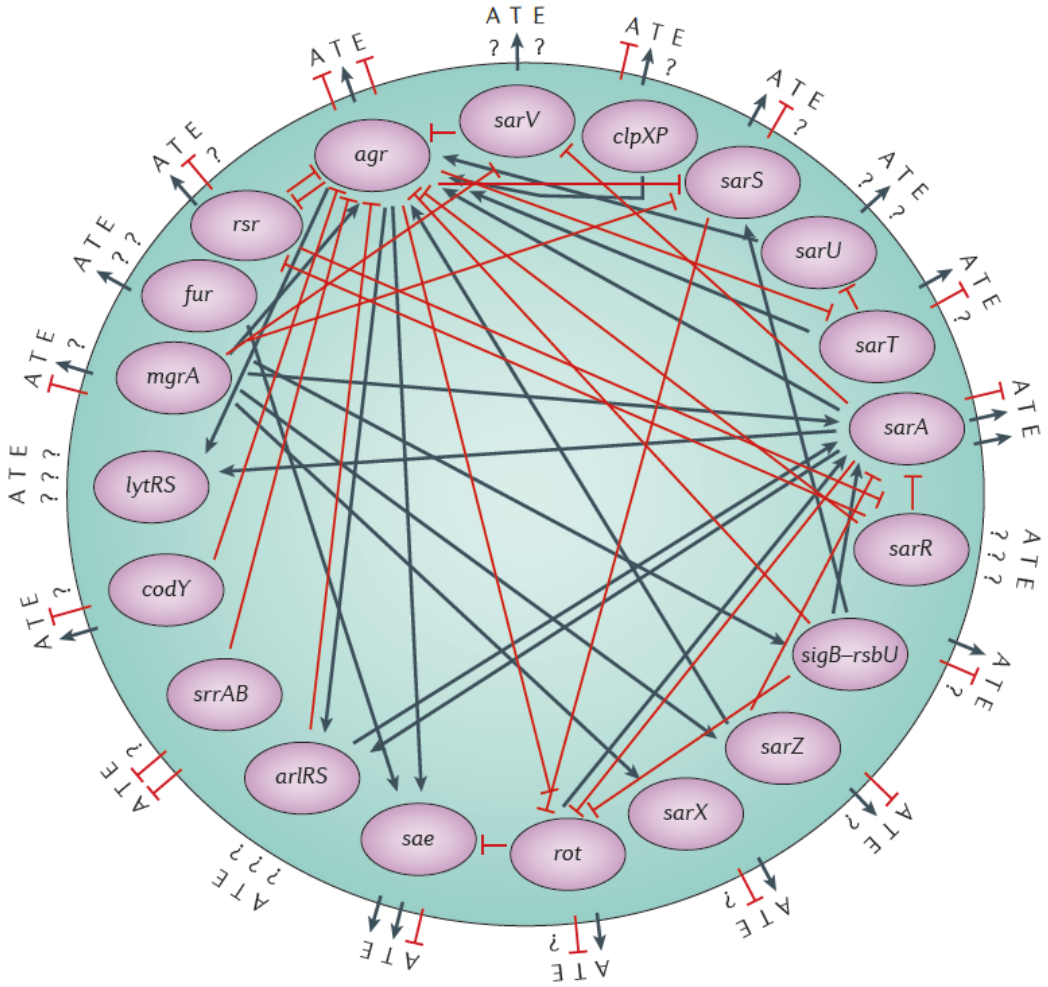


Figure 1.7. The known virulence regulatory network in *S. aureus*. This figure is taken from Priest et al. (2012), reused with permission from the corresponding author. Inside the circle are all the regulatory genes shown to affect each other and virulence. Outside the circle, the known effects of each regulator on adhesion (A), toxicity (T) and evasion (E) are indicated. Question marks indicate either no information is available regarding the direct activity of the regulator, or that available information is conflicting. The authors note that much of the data used to generate this image is qualitative.

Work by Peacock et al. (2002) further demonstrates the effect individual virulence-related genes can have in combination on the virulence phenotype of an isolate. Here, the authors studied the contributions of a variety of adhesins and toxins to invasive staphylococcal disease and found seven virulence factor genes which were significantly more common in invasive *S. aureus* isolates compared to carriage isolates. Interestingly, no single factor emerged as a predominant effector of virulence, but each appeared to have a cumulative effect on the virulent phenotype.

Production of virulence factors by *S. aureus* is affected by environmental signals and growth conditions, and changes in virulence factor expression in response to

these signals is predominantly regulated at the transcriptional level by two-component regulatory systems such as *agr*, *saeRS*, *srrAB*, *arlSR* and *lytRS* (Bronner et al., 2004). The most well-known of these is the global regulator *agr*.

1.3.3.1 The Agr quorum sensing system

The accessory gene regulator (*agr*) system is a two-component quorum sensing system and a global regulator of virulence in *S. aureus*. The operon consists of the genes *agrABCD*, two promoters P2 and P3 and the small regulatory RNA (srRNA) gene RNAIII, which also encodes the phenol soluble modulin δ-toxin (see figure 1.8).

AgrA is an intracellular regulatory protein which is phosphorylated by the membrane-bound histidine kinase AgrC (Wang et al., 2014; Rajasree et al., 2016). Phosphorylated AgrA then interacts with the divergent P2 and P3 promoters, stimulating expression of the *agr* operon via the P2 promoter, and expression of the global virulence regulator RNAIII via the P3 promoter (Koenig et al., 2004, Reyes et al., 2011). The first gene downstream of P2 is AgrB; a transmembrane protein that modifies the AgrD propeptide post-translationally and transports the modified AgrD, referred to as autoinducing peptide (AIP), through the membrane to the extracellular space (Ji et al., 1997; Otto et al., 1998; Zhang et al., 2002). AIP binds the protein histidine kinase receptor AgrC, stimulating autophosphorylation of the receptor which then transfers the phosphoryl group to AgrA (Lina et al., 1998; Wang et al., 2014), and the cycle continues.

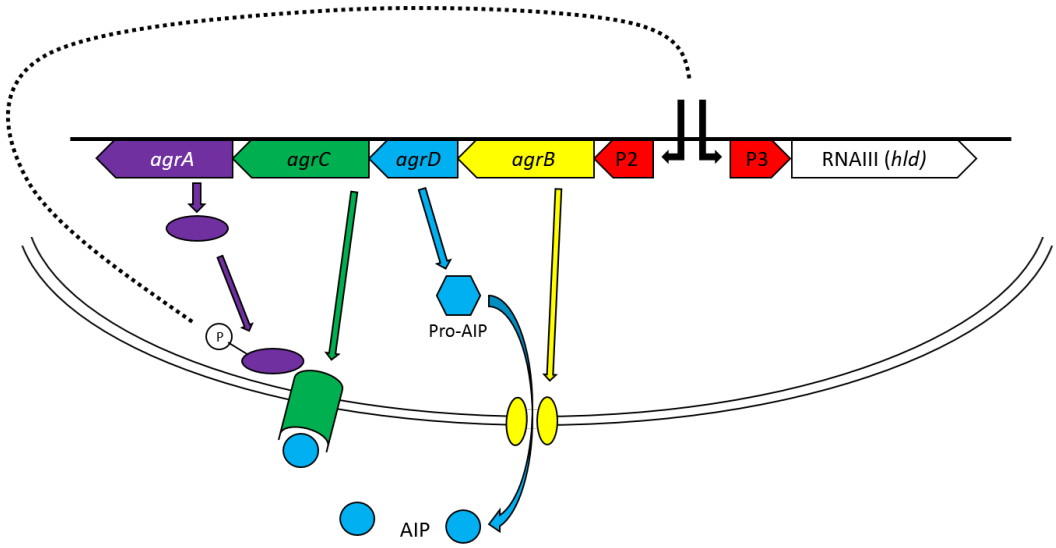


Figure 1.8. The *agr* operon in *S. aureus*. Diagram of the *agr* operon illustrating the role of each component in this two-component quorum sensing system.

From the P3 promoter, phosphorylated AgrA stimulates expression of the small regulatory RNAIII (sRNA), which regulates expression of several virulence genes; expression of adhesive proteins is downregulated, including those involved in biofilm production, as are immune evasion proteins such as protein A (Cheung et al., 1997; Vuong et al., 2000; Boles & Horswill, 2008), whilst expression of toxins and other virulence factors is simultaneously upregulated, for example α-, β- and δ-haemolysins, staphylokinase and nuclease (Recsei et al., 1986; Morfeldt et al., 1988). Adhesins and immune evasion proteins are downregulated at this stage as the bacterium transitions from colonisation of the host, for which both will be required, to invasion and persistence within the host, for which actively damaging toxic proteins are required.

RNAIII regulates gene expression by base pairing with target mRNA sequences to mediate translation, which is achieved through interactions with the Shine-Dalgarno sequence of the target mRNA (Gupta et al., 2015). This interaction can either lead to inhibition of translation in which RNAIII binds the Shine-Dalgarno sequence (as in the case of adhesive proteins) or facilitation of translation whereby RNAIII unmasks the Shine-Dalgarno sequence (as in the case of alpha-toxin). *S. aureus* mutants deficient in Agr activity have been shown on several occasions to be severely attenuated in virulence (Janzon & Arvidson, 1990; Abdelnour et al., 1993; Gillaspy et al., 1995), demonstrating the importance of this system in pathogenic *S. aureus* isolates.

As cell density increases, AIP levels increase concomitantly, stimulating further phosphorylation of AgrA in cells within the population, and in this manner creating a positive feedback loop. Toxin production and secretion therefore increases with Agr activity and decreases with its repression; effects on this system are usually the first avenue of investigation upon discovery of toxicity-deficient mutants in *S. aureus*.

1.3.3.2 The TCA cycle

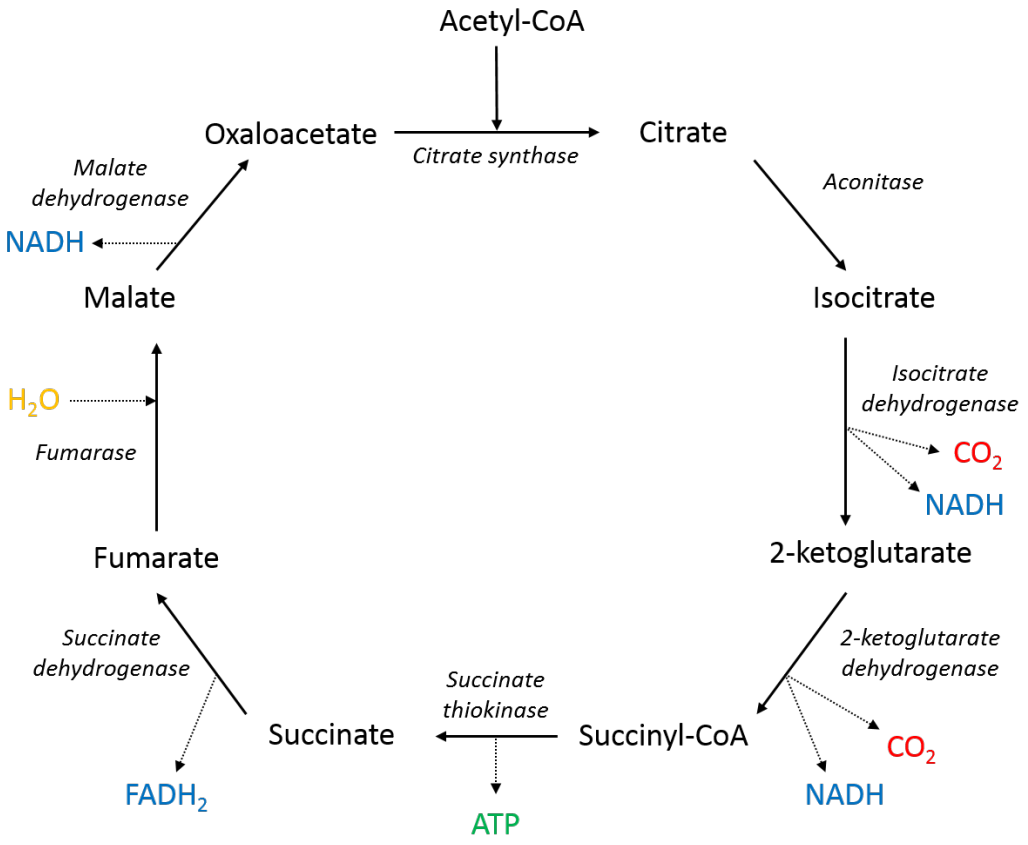


Figure 1.9. The Tricarboxylic Acid (TCA) Cycle in *S. aureus*. Diagram of the TCA cycle in *S. aureus* illustrating each intermediate involved in the cycle, the enzymes which catabolise each step, the energy, reduced dinucleotides and CO_2 that are produced and the H_2O that is used throughout the cycle.

The tricarboxylic acid (TCA) cycle is a key part of metabolism in *S. aureus*. During aerobic growth of *S. aureus in vitro*, in the presence of a catabolisable carbon source and glutamate, the TCA cycle is repressed throughout the exponential growth phase (Somerville et al., 2003a). Pyruvate is metabolised via glycolysis producing acetyl-coenzyme A, which is converted to acetylphosphate before being used to generate ATP (Somerville et al., 2002; Somerville et al., 2003a). A by-product of this reaction is acetate, which accumulates exogenously in the cell medium (Somerville et al., 2002). Once the bacteria enter stationary phase, and the supply of external amino acids becomes growth limiting (Zhu et al., 2009), the TCA cycle is derepressed such that the accumulated acetate in the culture medium is catabolised in place of pyruvate (Somerville et al., 2003a). This produces the necessary energy required for continued growth (see fig 1.10).

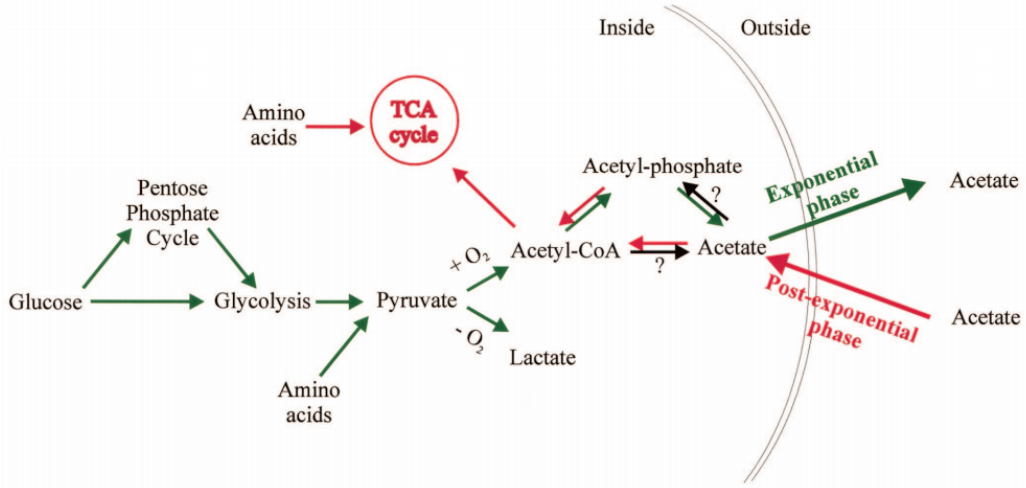


Figure 1.10. **Glucose catabolism by *S. aureus*.** Illustration of the reactions which take place during the exponential phase (green arrows) and reactions which take place during the post-exponential phase (red arrows) in *S. aureus*. Black arrows indicate reactions for which there are insufficient data to determine whether they occur. Figure from Somerville et al. (2003a) reproduced with permission, copyright American Society for Microbiology.

When the TCA cycle is active, reduced dinucleotides (eg. NADH, $FADH_2$) produced from the cycle donate electrons to membrane-bound electron acceptors in the electron transport chain, producing energy in the form of ATP, as shown in figure 1.11 (Dickerson et al., 1976; Schaetzle et al., 2008).

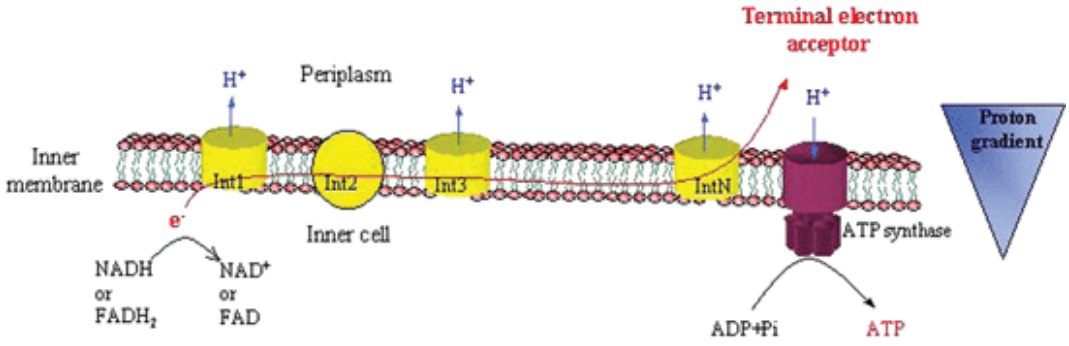


Figure 1.11. **Donations of electrons from reduced dinucleotides to the electron transport chain**. Reduced dinucleotides produced from the TCA cycle donate electrons to the electron transport chain, where they are used to fuel ATP production. Figure from Schaetzle et al., 2008, reproduced with permission from the Royal Society of Chemistry.

There is therefore a close link between electron transport and TCA cycle activity; in the absence of a functional electron transport chain, the TCA cycle can be repressed through a process of feedback inhibition, possibly mediated through the action of the transcriptional regulator ArcA. For example, in *E. coli* it has been shown that ArcA, part of a two-component system, represses expression of

genes involved in respiratory metabolism under anaerobic or microaerobic conditions, including enzymes of the TCA cycle (Georgellis et al., 2001; Malpica et al., 2004; Perrenoud & Sauer, 2005). ArcA also simultaneously upregulates expression of genes involved in fermentation, allowing energy production to continue in the absence of aerobic respiration (Georgellis et al., 2001).

1.3.3.3 Metabolism, Agr activity and toxin production

Changes in cell metabolism can have important consequences for production of virulence factors including toxins, which are energetically costly to produce, and mutants deficient in TCA cycle activity have been shown many times to exhibit attenuated virulence (Somerville et al., 2002; Somerville et al., 2003b; Zhu et al., 2009; Sadykov et al., 2010), as have mutants in electron transport (McNamara & Proctor, 2000; Kohler et al., 2003). Work by Proctor & Peters (1998) discusses how decreasing ATP levels can activate the stress sigma factor σ^B. This is thought to interact with the global regulator *sar*, which in turn interacts with the *agr* system, and in the absence of sufficient levels of ATP this chain of interactions results in downregulation of *agr* and a subsequent reduction in toxin production.

Evidence for this chain of interactions is corroborated in work by Chan & Foster (1998), who demonstrate that the global virulence regulator SarA responds to changes in aeration during growth to influence expression of the toxin α-haemolysin and a number of staphylococcal proteases. Further work by Chan et al. (1998) investigates the interactions between SarA and σ^B and demonstrates a close relationship between environmental, and therefore metabolic stresses, and subsequent regulation of virulence factor production. Cell metabolism is therefore tightly linked to pathogenicity in *S. aureus*.

1.3.5 Prior research into *S. aureus* toxicity

Interestingly, community-acquired isolates have been shown to be more toxic to the host than hospital-acquired isolates (Rudkin et al, 2012). This is thought to be because healthy people in the community will present more of a challenge to the pathogen, and as such it must be more virulent to colonise and cause disease in an otherwise healthy host. Whereas in patients who are already vulnerable to infection, even isolates of low virulence will be able to colonise the host and cause disease.

Following on from this, work by Laabei et al (2015) has shown that *S. aureus* isolates which cause severe bacteraemia (infection of the blood) are less toxic than isolates which cause relatively mild skin and soft tissue infections. This is thought to result from a trade-off between virulence and disease severity, because toxin production is an energetically expensive activity, and as the bloodstream is quite a hostile environment for the bacteria it acts as an extreme bottleneck. As such only the fittest bacteria (i.e. those not expending unnecessary energy on toxin production at this specific stage of disease) successfully survive this.

These results stemmed from a genome-wide association study (GWAS) conducted in *S. aureus* which aimed to identify novel toxicity-affecting loci that contribute to the wide-ranging toxic phenotypes observed in this bacterium. From this GWAS and following functional verification of the results to identify true-positive associations with the phenotype, numerous loci were identified as affecting toxicity. One of these, *cyoE*, is the subject of the first chapter of this thesis, which is characterised in-depth to identify the mechanism by which it exerts it's effect on toxicity in *S. aureus*.

Research such as this is important in making advances towards a better understanding of virulence in *S. aureus*, which in turn will contribute to the development of better treatments and more efficient use of the antibiotic treatments which are still effective. While there is specific focus on a single gene here, this kind of research is important if the function of individual genes in relation to virulence is to be understood in the context of the wider genetic network of toxicity effectors. Focussing on individual genes which are associated with the toxic phenotype also has the potential to identify novel drug targets, which in the short term may help to reduce the burden on current antibiotics as a first resort for treatment by expanding the current pool of available antibiotics.

A long-term aim of research such as this is to eventually develop a system of sequenced-based diagnosis, in which an infecting isolate can be sequenced quickly and accurately to determine its genetic profile, the specific treatment required by an individual patient, and any potential complications which may arise. In terms of moving towards this long-term goal, whole genome sequencing of bacterial isolates in real time during disease outbreaks has proven to be an extremely useful tool (eg. Quick et al., 2016), despite being somewhat limited to studying the evolution of bacterial pathogens and the epidemiology of disease

transmission in many cases (Young et al., 2012; Harris et al., 2010). However, simply knowing the genome sequence of an isolate is of little use unless we can fully understand the function of genes which contribute to the pathogenic phenotype. This point is highlighted by Viney (2014), which stresses the importance of understanding individual gene function in an era when it is becoming increasingly common to compare similarities between genomes instead of focussing on the function of specific genes. If whole genome sequencing is to be used in the future to improve disease treatment, it is necessary to study individual gene function in relation to pathogenesis so that pathogenic genomes can be sequenced and interpreted in a way which will be clinically useful.

1.4.0 Introduction to *Streptococcus pneumoniae*

1.4.1 Pneumococcal disease severity and prevalence

Streptococcus pneumoniae is an opportunistic pathogen which normally inhabits the nasopharynx in a small proportion of the population as a commensal, but can cause life-threatening invasive disease in the form of pneumonia if it is able to enter the respiratory tract, meningitis if it is able to breach the blood-brain barrier, and septicaemia if it is able to enter the bloodstream (Ioachimescu et al., 2004; O'Brien et al., 2009; Deng et al., 2016). Carriage rates vary, with children below the age of five having the highest carriage rates of up to around 50-60% (Regev-Yochay et al., 2003; Hussain et al., 2005; Henriques-Normark & Normark, 2010), while carriage rates in adults tend to be much lower at around 2-20% (Hussain et al., 2005; Hendley et al., 1975; Costello et al., 2016).

Young children are at a particularly high risk of serious invasive disease, with an estimated 800,000 deaths worldwide in children under five attributable to invasive pneumococcal disease each year (Deng et al., 2016). The burden of pneumococcal disease has been such that the World Health Organization recently announced a global action plan to help end preventable child deaths from pneumonia by 2025; this is hoped to be achieved through improving hygiene, and increasing availability and uptake of the pneumococcal vaccine in poor and remote communities where prevalence of the disease is high (WHO & UNICEF, 2013).

While access to vaccines and antibiotic treatment for pneumococcal disease is much better in developed countries, pneumococcal infections are becoming

increasingly problematic in patients with other underlying health conditions such as COPD or previous respiratory tract infection and in hospitalised patients with compromised immunity (Niederman & Zumla, 2016; Sibila et al., 2016). Complications also arise when infection is caused by multi-drug resistant strains of *S. pneumoniae*, which although not yet at epidemic levels, are also becoming an increasing problem.

A further issue with controlling the spread of disease caused by *S. pneumoniae* is that there are over 90 known serotypes of this bacterium (Henrichsen, 1995; Mehr & Wood, 2012), which differ based on the antigenic properties of their polysaccharide capsule. Vaccines currently in use cover between 7 and 23 of the most virulent serotypes, but the 23-valent vaccine is not immunogenic in children below 2 years old, and therefore cannot be administered to the most at-risk age group. To induce an immune response in children under 2, a 13-valent conjugate vaccine was developed in which capsular antigens are conjugated to inactivated diphtheria toxin.

While these pneumococcal vaccines confer protection against the serotypes they target, clinicians are beginning to see increasing evidence of serotype replacement, whereby non-vaccine serotypes of *S. pneumoniae* which have not previously caused high levels of disease become more prevalent in clinical samples obtained from pneumococcal infection (Spratt & Greenwood, 2000; Brueggemann et al., 2003), while serotypes included in the vaccine become less prevalent. This is illustrated in figures 1.12 and 1.13 below.

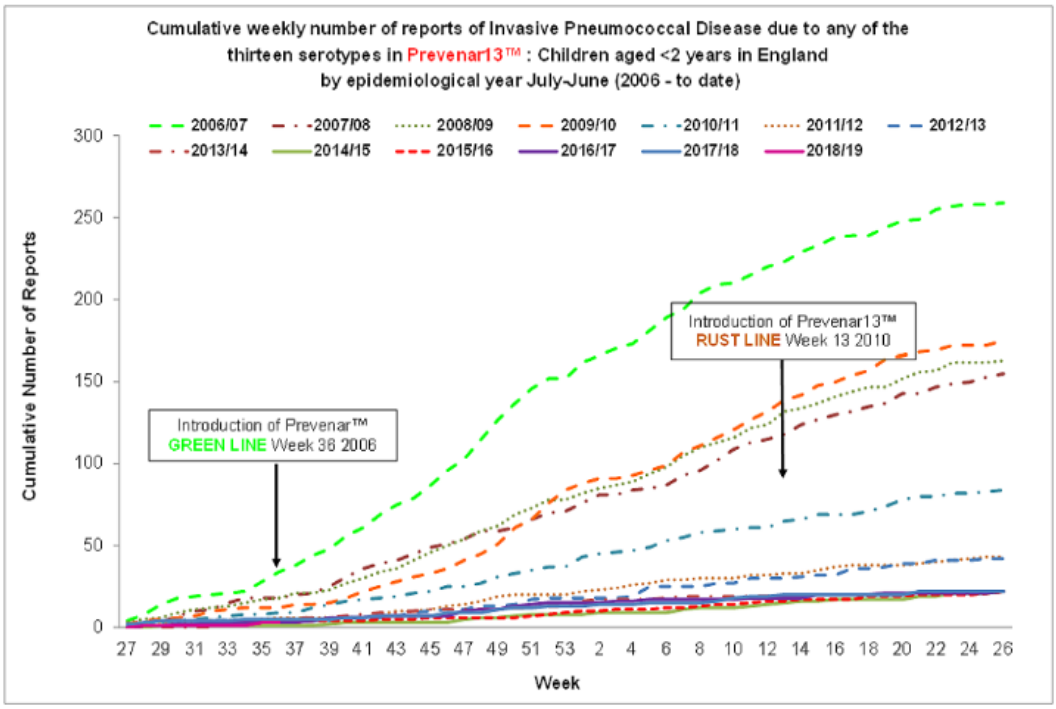


Figure 1.12. Incidence of invasive pneumococcal disease in children <2yrs caused by vaccine serotypes from 2006-2018/19. Disease caused by strains included in the Prevenar13 vaccine is shown to be reduced in incidence over the period from 2006-2019. Figure from Public Health England (2018b), reused under the Open Government Licence v3.0.

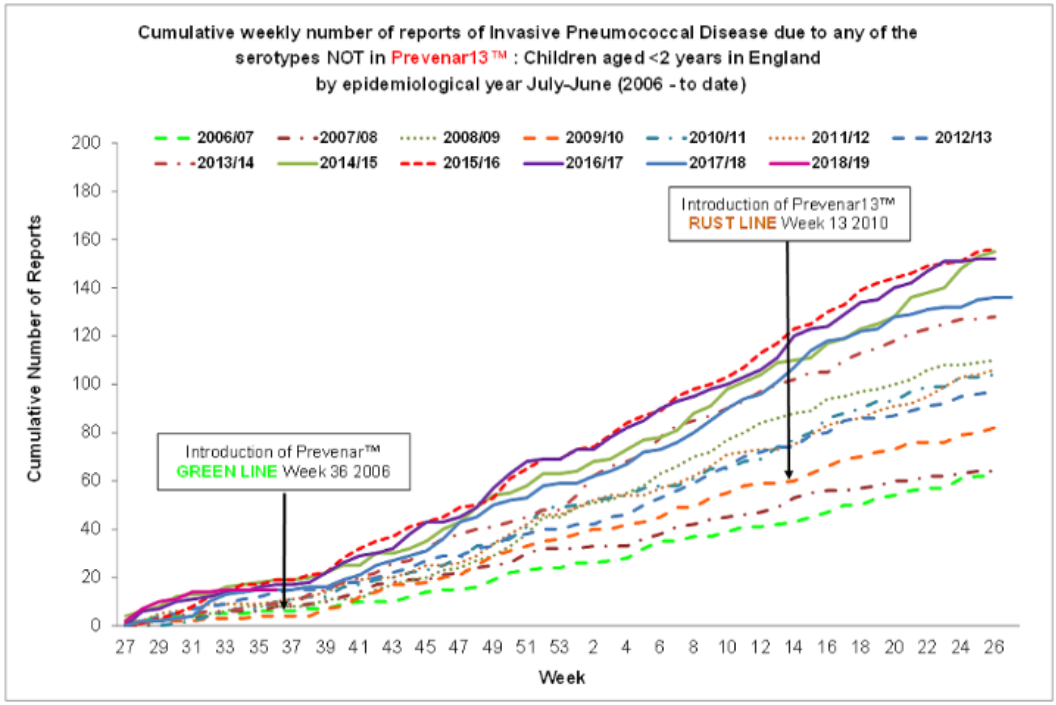


Figure 1.13. Incidence of invasive pneumococcal disease in children <2yrs caused by non-vaccine serotypes from 2006-2018/19. Disease caused by strains not included in Prevenar13 is shown to have increased in incidence since 2006, showing the inverse pattern to figure 1.12. Figure from Public Health England (2018c), reused under the Open Government Licence v3.0.

Clearly, there is a need to reduce the burden of disease caused by *S. pneumoniae*, and with increasing levels of antibiotic resistance being observed there will soon be an urgent need to find alternative treatments for this pathogen. To achieve this, we need to fully understand how this pathogen causes disease and in turn use this knowledge to find new targets for treatment of disease.

1.4.2 *Streptococcus pneumoniae* virulence and pathogenicity

Streptococcus pneumoniae has an array of virulence factors, including the antiphagocytic polysaccharide capsule, a wide range of adhesins and cell surface proteins, and the toxin pneumolysin. Interestingly, and unlike other opportunistic pathogens which inhabit the nasopharynx, pneumolysin is the only toxin produced by this bacterium. The range of virulence factors found in this pathogen and their locations within the cell are illustrated in figure 1.14 below, from van der Poll & Opal (2009).

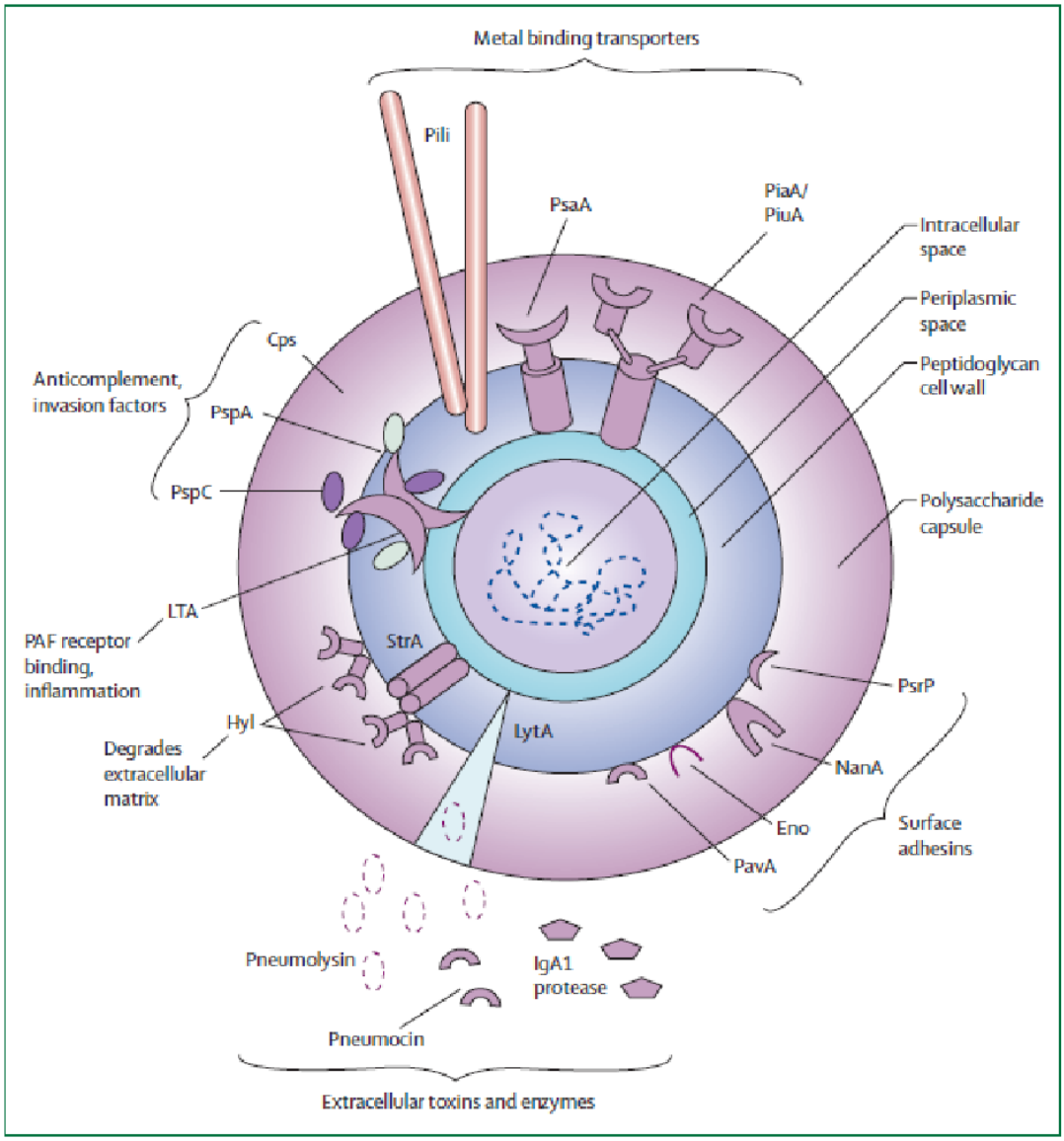


Figure 1.14: Virulence factors of pneumococcus. PsaA= pneumococcal surface antigen A. PiaA/PiuA= pneumococcal iron acquisition & uptake. PsrP=pneumococcal serine-rich repeat protein. NanA=neuraminidase. Eno=enolase. PavA=pneumococcal adhesion & virulence. LytA=autolysin. StrA=sortase A. Hyl=hyaluronate lyase. LTA=lipoteichoic acid. PspC=pneumococcal surface protein C. PspA=pneumococcal surface protein A. Cps=polysaccharide capsule. PAF=platelet-activating factor. Figure from van der Poll & Opal (2009), reproduced with permission from Elsevier.

1.4.2.1 The pneumococcal capsule

The polysaccharide capsule is a key virulence factor for this pathogen, and antigenic differences between encapsulated strains are responsible for the large number of different serotypes within the species which is such a problem for vaccine development. The capsule has several virulence-related functions including antiphagocytic activity, prevention of removal of the pneumococcus by

mucus in the respiratory tract and reducing sensitivity to antibiotics (Mitchell & Mitchell, 2010).

Several groups have investigated the role of the pneumococcal capsule in colonisation of the upper respiratory tract during the initial stages of infection. Kim & Weiser (1998) explored the differences in capsule production between the opaque and transparent colony phenotypes of *S. pneumoniae* and the resulting effect on virulence. It was shown that the more virulent opaque phenotype was generally associated with greater amounts of capsular polysaccharide and systemic infection, while the transparent phenotype with comparatively reduced capsular polysaccharide in general was selected for during colonisation. However, Magee & Yother (2001) observed that the ability of capsule specific antibodies to be protective against pneumococcal carriage, in combination with variations in the ability of different serotypes to colonise the nasopharynx based on capsular differences, suggests an important role for the capsule during the colonisation process. The authors go on to demonstrate that capsular polysaccharide is essential for colonisation by strains of serotypes 2 and 3, with unencapsulated strains shown to be unable to colonise the nasopharynx in a mouse model of infection.

It is well established that the capsule is essential for virulence during the later stages of pneumococcal disease, since a body of work has shown that unencapsulated strains suffer a significant loss in virulence compared to wild type encapsulated strains (Avery & Dubos, 1931; Watson & Musher, 1990; Bender et al., 2003). The specific mechanism by which the capsule contributes to virulence was confirmed in work by Hyams et al. (2010), in which the authors showed that the capsule prevents opsonisation of the bacterium with factors C3b and iC3b of the alternative and classical complement pathways, and inhibits conversion of surface-bound C3b to iC3b. Binding of IgG and C-reactive protein to unencapsulated pneumococci suggested that the capsule also serves to mask antibody recognition of subcapsular antigens to inhibit opsonophagocytosis and it was further shown that the capsule inhibits a multitude of receptors to interfere with neutrophil phagocytosis.

Work by Wartha et al. (2007) additionally confirms the role of the capsule in interference with host neutrophils; here the authors show that presence of a capsule around the pneumococcal cell can protect against the action of neutrophil extracellular traps. Specifically, it was shown that capsule expression

reduces the extent to which pneumococcal cells remain trapped, and that in combination with the incorporation of D-alanine to lipoteichoic acid on the bacterial membrane, the capsule facilitates resistance to the action of antimicrobial peptides (AMPs) produced by neutrophils by creating a positive charge that repels the likewise positively charged AMPs.

Interestingly, pneumococci were shown in work by Hammerschmidt et al. (2005) to modulate capsule expression at different stages of infection, such that pneumococcal cells in contact with host cells express a reduced amount of capsule while those not in contact with the host exhibit a thick capsule layer. It was demonstrated that this resulted from active down-regulation of capsule as opposed to the enrichment of acapsular mutants within the population. This reduction of the capsular polysaccharide layer facilitates adhesion and invasion of the pneumococcus, exposing surface proteins and adhesins within the cell wall which promote attachment to host epithelial cells. However, capsule production must be strictly regulated to prevent exposure of the pneumococcus to clearance by the host immune system.

1.4.2.2 Surface proteins and adhesins

Several virulence related surface proteins have been described in the literature which contribute to pathogenesis throughout the stages of colonisation and invasion. These include LPXTG-anchored surface proteins, lipoproteins, and choline-binding proteins. Figure 1.15 depicts some of the key families of pneumococcal surface proteins.

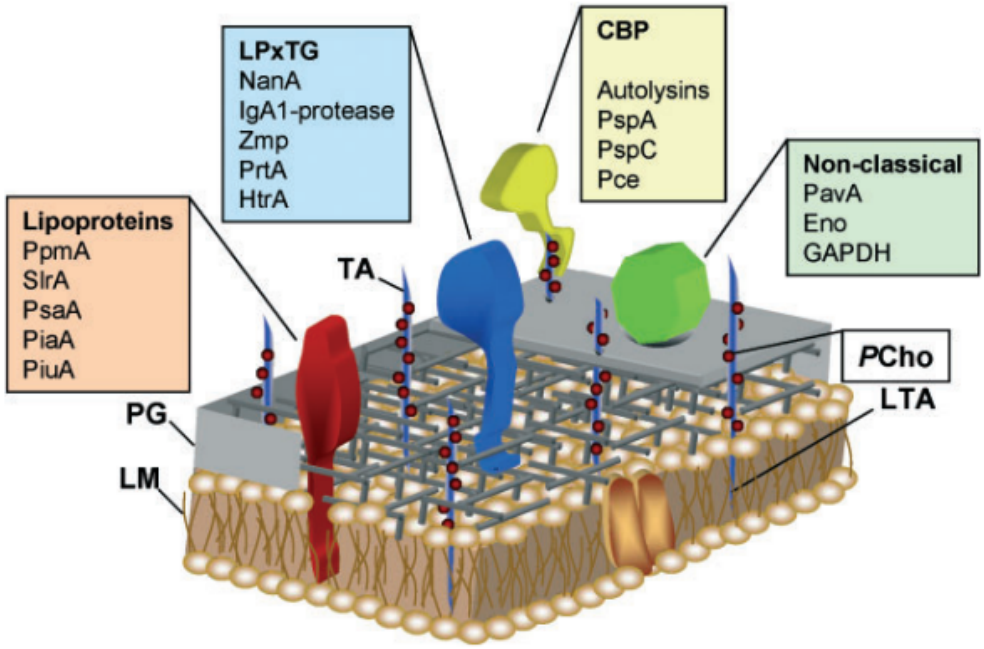


Figure 1.15. The pneumococcal cell wall and surface-exposed proteins. The pneumococcal cell wall consists of a phospholipid membrane (LM), peptidoglycan (PG), and teichoic and lipoteichoic acids. The cell wall also contains phosphorylcholine (*P*Cho), which anchors choline-binding proteins (CBPs) non-covalently on the cell wall. Virulence proteins of the different classes of pneumococcal surface proteins are depicted. From Bergmann & Hammerschmidt (2006). Reproduced with permission from the Microbiology Society.

LPXTG-anchored proteins

Table 1.1. Names and functions of pneumococcal LPXTG-anchored surface proteins.

Protein name	Function
NanA (neuraminidase)	Cleaves terminal sialic acid residues on cell surfaces
IgA1-protease	Cleaves the hinge region of human IgA1
Zmp	Zinc metalloprotease involved in tissue disruption and invasion
PrtA	Protease with unknown function
HtrA	Post-translational regulator involved in the stress response, competence and cell division

LPXTG-anchored surface proteins are so named because of the LPXTG amino acid motif found at the C-terminus of these proteins that acts as a substrate for sortase enzymes during the anchoring of proteins to the cell wall (Navarre & Schneewind, 1994). This group encompasses several proteases, including the enzymes IgA1-protease, neuraminidase (NanA) and the serine proteases HtrA

and PrtA. These enzymes act to break down connective tissue, extracellular matrix and cell wall components within the host, facilitating invasion of host tissues and establishment of infection (Mitchell & Mitchell, 2010).

IgA1-protease acts to cleave the hinge region of human IgA1 bound to the bacterial cell surface, thereby subverting the immune response and aiding persistence within the host (Chi et al., 2016). NanA cleaves terminal sialic acid residues on cell surfaces, affecting the function of immune defence proteins such as lactoferrin and IgA2 (King et al., 2004), modifying the surface of competing bacteria to gain an advantage within the niche, and exposing potential receptors on the epithelial cell surface to which the pneumococcal cell can then adhere (King et al., 2006; Parker et al., 2009). NanA has been shown to play a major role in adherence, colonisation and to a lesser extent in biofilm formation during disease to promote survival and persistence of the pneumococcus within the host (Brittan et al., 2012).

The serine proteases HtrA (high temperature requirement A) and PrtA (cell wall-associated serine protease A) have both been shown to contribute to virulence in models of pneumococcal disease. HtrA is a posttranslational regulator which functions in the stress response, competence and cell division in the pneumococci, and HtrA-deficient mutants exhibit attenuated virulence (Ibrahim et al., 2004; de Stoppelaar et al., 2013). Likewise PrtA has been shown to play an important role in intraperitoneal infection and lung inflammation, with PrtA-deficient mutants exhibiting reduced virulence (Bethe et al., 2001; Hsu et al., 2018). However, the exact role of this protease is unknown (Mitchell & Mitchell, 2010).

Lipoproteins

Table 1.2. Names and functions of pneumococcal lipoproteins.

Protein name	Function
PsaA	Permease involved in adhesion to host epithelial cells
PiaA	Involved in iron acquisition and highly immunogenic
PiuA	Involved in iron uptake and highly immunogenic
PlpA	Binds to glycoconjugate receptors on pneumocytes to facilitate colonisation
Ami	Binds to glycoconjugate receptors on pneumocytes to facilitate colonisation

Lipoproteins such as pneumococcal surface adhesin A (PsaA) play an important role in adhesion to host epithelial cells during colonisation. PsaA is an ABC-type permease complex involved in the transport of manganese ions across the cell membrane and adhesion to host epithelium (Rajam et al., 2008). Anderton et al. (2007) demonstrated that E-cadherin, the eukaryotic cell-cell junction adhesin, is the receptor for PsaA in host epithelial tissue. When the function of PsaA is disrupted, a loss in virulence and adhesive ability is observed, along with increased sensitivity to oxidative stress (Berry & Paton, 1996; Dintilhac et al., 1997). Interestingly, Ogunniyi et al. (2002) found PsaA to be one of the most abundant virulence proteins at various time points after intraperitoneal infection, suggesting an extensive role in pneumococcal virulence for this protein.

Other pneumococcal ABC transporters, including PiaA (pneumococcal iron acquisition protein A) and PiuA (pneumococcal iron uptake protein A), have been shown to be important for pneumococcal virulence in systemic and pulmonary murine models of infection (Brown et al., 2001), and elicit a strong, serotype-independent immunological response in infants as young as 7 months old (Whalan et al., 2005). Peptide permeases such as PlpA and Ami have also been found to play an important role in colonization of the respiratory tract by binding to glycoconjugate receptors on the surface of pneumocytes - the cells that line the walls of the alveoli (Cundell et al., 1995; Salyers & Whitt, 2002).

Choline-binding proteins

Table 1.3 Names and functions of pneumococcal choline-binding proteins.

Protein name	Function
CbpA/PspC	Adhesion to host epithelial cells
LytA	Autolysis of cells within the population
PspA	Inhibits opsonization and binds lactoferrin

The choline-binding protein family are another group of important contributors to pneumococcal virulence, and similarly to PlpA and Ami, proteins such as choline-binding protein A (CbpA) have been shown to adhere to type II pneumocytes and glycoconjugate receptors on host epithelial cells. When disrupted in animal models, pneumococci deficient in CbpA exhibit a loss in ability to colonise the nasopharynx (Rosenow et al., 1997; Paton, 1998). This family of proteins also includes one of the major contributors to virulence in pneumococci; the autolytic

enzyme LytA which causes lysis of a proportion of the population in the stationary phase of growth. This phenomenon is distinguishable from apoptosis in that it can be induced in one cell through the action of a related, predatory cell within the population. Apoptosis on the other hand is defined as a form of 'programmed' cell death, which is rarely discussed in the literature in the context of the pneumococcus as the focus is usually on virulence-related phenotypes. Autolysis is discussed in further detail in section 1.4.2.3 below.

Non-classical proteins

Table 1.4 Names and functions of non-classical pneumococcal surface proteins.

Protein name	Function
PavA	Modulates adhesion to and invasion of host tissues by binding to fibronectin
Eno (enolase)	Binds extracellular matrix proteins
GAPDH	Role in invasion and modulation of host immune system
Pilus	Adhesion to cells of respiratory tract
Phosphorylcholine	Anchors choline-binding proteins to cell wall

Other surface proteins and adhesive virulence factors which do not fit into one of the above-mentioned categories include the pneumococcal pilus, which is found in a small number of clinical strains (Barocchi et al., 2005), and phosphorylcholine, which anchors the choline-binding proteins to the cell wall (Bergmann & Hammerschmidt, 2006). Analysis of the *S. pneumoniae* genome by Bagnoli et al. (2008) highlighted the presence of genes which encode two types of pilus islets, thought to play an important role in initial attachment and colonisation of the respiratory tract and in eliciting an inflammatory response within the host. Cell wall phosphorylcholine has been shown to bind host cell platelet activating factor (PAF) to promote inflammation (Cundell, Masure & Tuomanen, 1995; Jenkinson, 1997).

The surface proteins and adhesins discussed here are just some examples of the extensive range of proteins *S. pneumoniae* employs to colonise and invade the host respiratory tract. One of the most interesting and perplexing of these is the autolytic enzyme LytA, which has been extensively studied for its contribution to virulence and its link to competence in this bacterium.

1.4.2.3 Autolysis

An interesting characteristic of the pneumococcus is its propensity to undergo autolysis in the stationary phase of growth. Autolysis is the destruction of cells within a population as a direct result of the action of their own lytic enzyme(s). When the pneumococcal population reaches a particular density, typically during the stationary phase of the life cycle, a proportion of cells within the population will lyse in response to the action of autolysin, a single lytic enzyme encoded by the gene *lytA*. This results in the release of cytoplasmic components into the extracellular environment, including pneumolysin, inflammatory cell wall fragments, a host of virulence factors, and non-virulent components including nutrients and DNA.

Many suggestions have been put forward in the literature as to why this process occurs in pneumococci, and it is still largely unclear what advantage there is to be gained from a bacterial population inducing lysis of a proportion of its own cells. Possible explanations include:

- To release pneumolysin and other virulence factors for the purpose of acquiring nutrients from host cells during disease (Berry et al., 1989; Canvin et al., 1995; Hirst et al., 2008)
- Fratricidal lysis whereby cells within the population which are not competent for genetic transformation are purposefully lysed by neighbouring competent cells (Claverys & Håvarstein, 2007; Guiral et al., 2005) – discussed further in section 1.4.2.5 below
- To release intracellular components for evasion of the host immune system during disease (Martner et al., 2009; Mellroth et al., 2012)
- To separate daughter cells at the end of cell division (Ronda et al., 1987)

Autolysin is an N-acetylmuramyl L-alanine amidase (Howard & Gooder, 1974), and a member of the choline-binding family of proteins. It acts specifically to cleave the bond between muramic acid and alanine in the peptidoglycan cell wall (Howard & Gooder, 1974; Mosser & Tomasz, 1970), causing the cell wall to break apart and release intracellular components into the extracellular environment. Much research into autolysin has focused on its potential fundamental role in pneumococcal virulence, based primarily on observations that autolysin-deficient mutants show reduced virulence in comparison to wild type strains (Berry et al., 1989; Canvin et al., 1995; Hirst et al., 2008).

Berry et al. in 1989 presented an early example of an autolysin-deficient mutant in strain D39 which showed a significant reduction in virulence *in vivo* compared to the wild type. Furthermore, Canvin et al. (1995) noted that unencapsulated pneumococcus was causing disease in agranulocytic animals (lacking neutrophils), while capsular polysaccharide was found to be non-inflammatory in these animals. Traditionally the polysaccharide capsule has been considered the most important virulence factor employed by the pneumococcus, with unencapsulated strains showing reduced virulence in comparison to encapsulated strains (Avery & Dubos, 1931; Watson & Musher, 1990); this work, however, demonstrated that the capsule was not the only key virulence factor in facilitating the onset of disease, and further to the work of Berry et al., it was demonstrated by Canvin et al. that mice infected with an autolysin-deficient strain of D39 showed no signs of disease, and notably the mutant was found to die *in vivo*.

The most obvious reason for the loss of virulence in these mutants would be that they no longer release virulence factors into the extracellular environment as a direct result of the deficiency in autolysin. Pneumolysin, for instance, lacks an N-terminal secretory signal domain (Price & Camilli, 2009), and therefore is not secreted by infecting cells, but instead is reliant on the action of autolysin for deployment into the extracellular environment. Likewise release of degraded peptidoglycan fragments and teichoic acid from lysed pneumococcal cells elicits a strong inflammatory response in an infected host as a direct result of the action of autolysin (Mitchell, 2000).

Interestingly, recent work looking at the relationship between the genes encoding autolysin and pneumolysin suggests that the *lytA* and *ply* genes are tightly linked on a pathogenicity island (Morales et al., 2015), and suggests a close evolutionary relationship between the two proteins which may add weight to the original theory that the primary role of autolysin is to do with the release of pneumolysin during infection.

1.4.2.4 Pneumolysin

The role of the toxin pneumolysin however is much clearer; pneumolysin is a major virulence factor for *S. pneumoniae* that is present in all serotypes (Hirst et al., 2004), and contributes greatly to the symptoms of disease observed in patients suffering from pneumococcal infection. Pneumolysin is a cholesterol-dependent cytolytic pore-forming toxin that is localised to the bacterial cell wall

(Price & Camilli, 2009). When released from cells, generally thought to be as a result of autolysis, it binds to cholesterol in the membranes of host epithelial cells, inserts into the lipid bilayer, and forms a transmembrane pore which ultimately causes lysis of the host cell (Paton, 1996). Depending on the location of infection, pneumolysin can cause extensive damage to the lungs and central nervous system. Several studies have illustrated significantly decreased virulence in animal models of infection when *S. pneumoniae* is deficient in this toxin, and it has long been accepted as a key virulence factor alongside the capsule (Canvin et al., 1995; Wellmer et al., 2002; Hirst et al., 2004; Hirst et al., 2008). Nonetheless, the contribution of pneumolysin to virulence extends much further than pore-formation and lysis of host cells (summarised in figure 1.16 below).

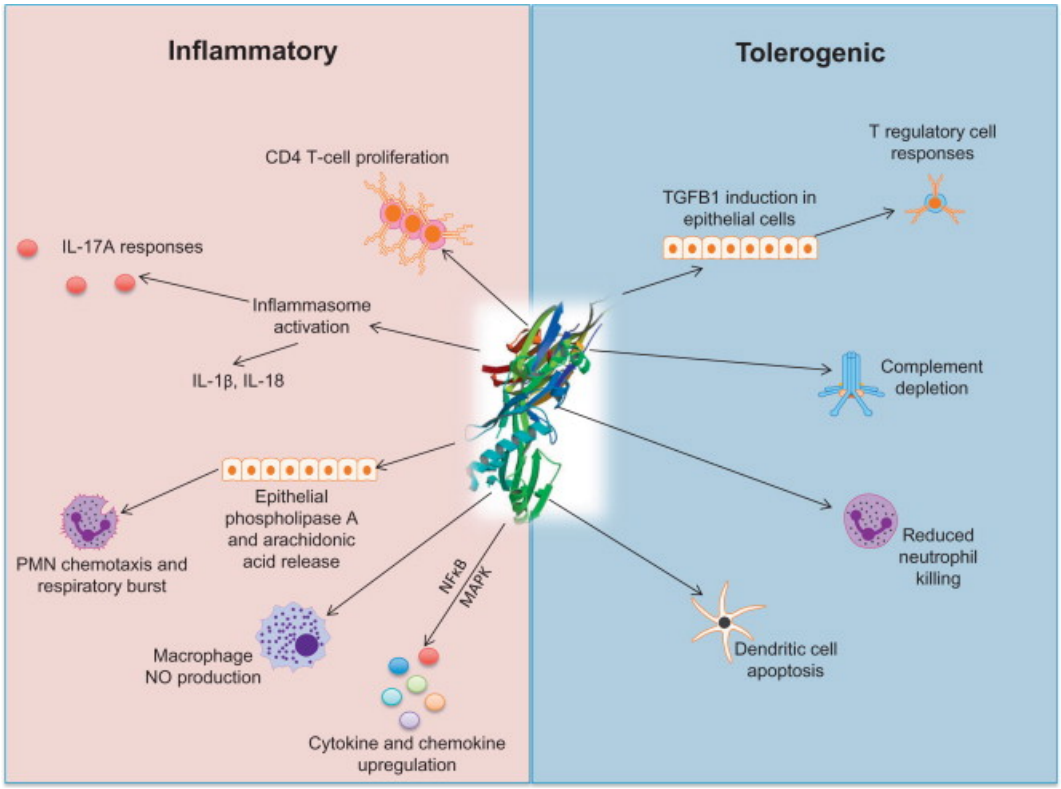


Figure 1.16. Inflammatory vs. tolerogenic properties of pneumolysin. The contribution of pneumolysin to virulence in *S. pneumoniae* is extensive, encompassing a variety of both inflammatory and tolerogenic properties which are used for different purposes depending on the stage of disease. From Neill et al. (2015), reproduced with permission from Elsevier.

Pneumolysin has been shown to activate complement; the classical complement pathway is activated by binding to C1q and the Fc region of human IgG (Mitchell et al., 1991), which increases inflammation and tissue damage in the host through over-production of the anaphylatoxins C3a and C5a (Paton et al., 1984). In doing so, pneumolysin also causes a depletion of complement components to limit the amount of complement available to opsonise pneumococcal cells; this

reduces complement-mediated clearance, which is an important aspect of host immunity to pneumococcal infection (Andre et al., 2017). In conjunction with pneumolysin's haemolytic activity on neutrophil cells, opsonophagocytosis is inhibited by the over-stimulation of complement, promoting evasion of the host immune system and persistence of pneumococci within the host (Rubins et al., 1995; 1996).

Further key interactions with host immunity have been demonstrated for pneumolysin, including proinflammatory interactions with Toll-like receptors (Malley et al., 2003), stimulation of a range of cytokines (Houldsworth et al., 1994; Cockeran et al., 2002; Shoma et al., 2008) and the NLRP3 inflammasome (McNeela et al., 2010; Witzenrath et al., 2011), and direct interference with the function of immune cells such as dendritic cells, macrophages and neutrophils (Littman et al., 2009; Bewley et al., 2014; Nel et al., 2016b).

Malley et al. (2003) demonstrated a specific relationship between pneumolysin and Toll-like receptor 4 (TLR4) produced by macrophages, which is responsible for the robust proinflammatory response of the host immune system to pneumolysin. The authors demonstrated that TLR4-deficient mice were significantly more susceptible to invasive pneumococcal disease than an isogenic wild type, and that the TLR4 response was independent of the cytolytic pore-forming function of pneumolysin, as demonstrated using a non-cytolytic mutant of the toxin which was found to stimulate TLR4 production similarly to the wild type. The result of TLR4 stimulation in the host is the activation of caspase-1 and subsequent stimulation of inflammatory cytokines, which has been shown to be partly dependent on the NLRP3 inflammasome (Fang et al., 2011).

The NLRP3 inflammasome plays an important role in immune clearance of pneumococci. NLRP3 stands for nucleotide-binding oligomerisation domain-like receptor family, pyrin domain containing 3 inflammasome (Witzenrath et al., 2011); it is a large multimeric protein complex responsible for activation of the proinflammatory cytokines interleukin-1β (IL-1β) and interleukin-18 (IL-18) via caspase-1, and initiates the process of pyroptosis, a form of programmed cell death which is crucial for controlling microbial infections (Bergsbaken et al., 2009; Yang et al., 2019). IL-1β and IL-18 regulate the response of the T-helper cells Th17 and Th1, respectively. In turn this determines the recruitment of neutrophils and macrophages to the site of infection and ultimately leads to bacterial clearance (van der Veerdonk et al., 2011). McNeela et al. (2010) identified

pneumolysin as an activator of the NLRP3 inflammasome and demonstrated that NLRP3 was required by the host for protective immunity against pneumococcal infection.

Witzenrath et al. (2011) expanded on this to demonstrate that NLRP3 was specifically activated by pneumococcal serotypes expressing a haemolytic form of the toxin, while those expressing variants of pneumolysin with reduced haemolytic activity, such as serotypes 1 and 8, did not activate NLRP3. Serotypes 1 and 8 are associated with lower case fatality rates and are less damaging to the host, however, their ability to better evade recognition by the host immune system, including the NLRP3 inflammasome, contributes to their increased prevalence in causing invasive disease such as parapneumonic effusion and empyema.

Of note, the genetic changes in the pneumolysin gene of the serotypes used by Witzenrath et al. provides an example of how population stratification and data filtration can affect statistical associations in GWAS; variations in this gene and their subsequent effect on bacterial virulence may be identified from GWAS approaches that account for a subset of different serotypes within a clonal complex, while approaches that do not account for population structure may overlook these associations because the genetic variants will be relatively rare within the population as a whole. Furthermore, the pneumolysin gene is generally highly conserved among pneumococcal strains, and so SNPs in this gene may be excluded from the initial data filtration step due to their relative rarity within a clonal complex.

On the other hand, over-stimulation of NLRP3 by pneumolysin leads to increased tissue damage and disease severity in the host. For example, work by Hoegen et al. (2011) demonstrated that activation of NLRP3 by pneumolysin contributed to increased brain injury during pneumococcal meningitis in a murine model, and mutants lacking inflammasome components responsible for caspase-1 activation had lower disease severity. Clearly therefore, activation of NLRP3 is a crucial aspect of host immunity required for effective clearance of a pneumococcal infection, however activity of the inflammasome can be manipulated by pneumolysin for the benefit of the bacterium and there is a fine line between an appropriate response to infection and a detrimental over-stimulation of host immunity.

The role of pneumolysin in disease extends yet further, with studies in recent years elucidating a role for the toxin in: inhibition of host defence cells such as microglia during meningitis (Hupp et al., 2019); induction of DNA damage and arrest of the cell cycle of host cells (Rai et al., 2016); promotion of host-to-host transmission of pneumococcal cells (Zafar et al., 2017); activation of platelets (Nel et al., 2016a) and stimulation of a wide range of host proinflammatory factors including TNF-α, interleukin-1 and neutrophil extracellular traps (Houldsworth et al., 1994; Nel et al., 2016b).

Recent work has also implicated pneumolysin in initial colonization of the nasopharynx, where the bacterium is able to reside asymptomatically in healthy individuals; Hotomi et al. (2016) discuss pneumolysin-dependent invasion of nasal mucosa in the initial establishment of colonization, suggesting that the toxin may also have an important role in behaviour which is not directly related to virulence.

A key problem in the treatment of invasive pneumococcal diseases is that the antibiotics which are most effective against *S. pneumoniae* infection are bacteriolytic and induce lysis of the bacterial cells (Hirst et al., 2004). As with the natural process of autolysis, this initiates the release of pneumolysin, inflammatory cell wall fragments and many other virulence factors which continue to damage host tissues in the short term, causing the symptoms of disease to worsen (McCullers & English, 2008; Nau & Eiffert, 2002; 2005). A better understanding of the genetic basis for pneumolysin production and its relationship to autolysis may prove useful in improving treatments for those suffering from serious invasive pneumococcal disease.

1.4.2.5 Competence

Autolytic activity and toxicity have been shown to be tightly linked in pneumococcal virulence. A third key component which contributes to virulence is competence for genetic transformation. *S. pneumoniae* was used in the initial experiments by Avery et al. (1944) to demonstrate the principle of genetic transformation and has since been used as a model organism for this process based on its characteristic of being naturally competent for uptake of DNA. Recent work has delved further into the role of competence in this species and has demonstrated an important link to both autolysis and virulence.

Competence - the state of being able to take up DNA from the extracellular environment - has been shown to directly influence lysis of cells within the pneumococcal population, whereby competent cells stimulate autolysis in non-competent, but otherwise identical cells. This has been termed ‘fratricidal lysis’ (Claverys & Håvarstein, 2007), in which the only determinant of which cells lyse and which do not is the expression of competence-related genes. The mechanism of fratricidal lysis is discussed by Guiral et al. (2005) in their investigation of a two-peptide bacteriocin system which they named the competence induced bacteriocin system CibABC, expression of which is induced upon entering the competent state. The authors propose that competent cells induce the lysis of non-competent cells through cell-to-cell contacts, in which the CibA and CibB molecules of competent cells interact with the surface of non-competent cells to trigger the action of LytA (along with two other hydrolases LytC and CbpD). Competent cells remain protected from the action of CibAB by producing an immunity factor, named CibC. While the exact mechanism of action of the CibAB complex is not known, it was speculated that in interacting with the surface of non-competent cells, CibAB may insert into these cells to disrupt the membrane structure and in turn deplete energy levels within the cell, leading ultimately to lysis by hydrolytic enzymes. Figure 1.16 illustrates the proposed CibAB mechanism of fratricidal lysis, termed ‘allolysis’ by Guiral et al..

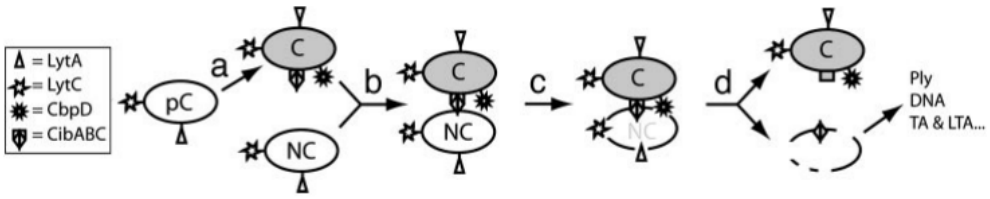


Figure 1.17. Proposed model of allolysis in *S. pneumoniae*. (A) Pre-competent (pC) cells differ from competent (C) cells by the presence of the amidase CbpD, the two-peptide bacteriocin CibAB and its immunity factor CibC. (B) Cell-to-cell contacts between competent and non-competent (NC) cells allows interaction of CibAB with NC cells, while competent cells are protected by CibC. (C) CibAB then triggers the action of CbpD, LytA and LytC on the cell wall of NC cells. (D) Cell wall disruption results in the release of intracellular components, including Ply, teichoic acid (TA), lipoteichoic acid (LTA) and chromosomal DNA. Figure from Guiral et al. (2005), copyright National Academy of Sciences.

Cell-to-cell contacts were established as necessary for this interaction, and this is supported in previous work by Steinmoen et al. (2003). Here, it was established

that the lytic enzymes produced by competent pneumococcal cells are not released into the growth medium but are likely to be anchored to the cell surface through choline binding domains. To induce lysis in non-competent cells therefore, cell-to-cell contact would be required.

'Allolysis', as defined by Guiral et al. (2005), differs from autolysis in that it is specifically triggered by the action of competent cells against non-competent cells within the population. This is not a suicidal response to nutrient depletion or environmental stress but is a predatory behaviour by a proportion of the population. Several possible reasons for the evolution of allolysis include the facilitation of genetic exchange, predation and competition among strains within a stationary phase population, or to facilitate the release of virulence factors from seemingly non-competent cells during infection.

The relationship between competence and virulence has been observed in earlier work by Lau et al. (2001) and Bartilson et al. (2001), where mutants with disruptions in competence-related genes have been found to have attenuated virulence in mouse models of infection. For example, Lau et al. (2001) observed that mutants in genes of the *com* regulon, specifically *comB* and *comD*, were found to be attenuated in a bacteraemia model and in both bacteraemia and pulmonary models of infection respectively. ComB is a component of the ComAB ABC transporter system which is involved in processing and secretion of the autoinducing competence stimulating peptide (CSP), while ComD is part of the ComDE two-component regulatory system involved in regulating the competent state of the pneumococcus in response to CSP. A mutation in the *comD* gene was also found to attenuate virulence in both systemic and lung colonisation infection models by Bartilson et al. (2001). Competence is a density-dependent phenotype; as cell density increases, CSP accumulates in the external environment, inducing competence in increasing numbers of cells proportional to the accumulation of CSP.

1.4.2.6 Competence, autolysis and virulence

Lau et al. and Bartilson et al. speculate that the induction of competence may be a response to environmental stresses that the population may encounter during stationary growth. Claverys et al. (2006) and Lin et al. (2016) expand further on this hypothesis by discussing the role of the regulator *comX* in induction of the competent state. ComX is an alternative σ factor which up-regulates a number of genes in response to the detection of CSP. However,

not all the genes up-regulated by ComX are required for DNA uptake, suggesting that induction of competence not only facilitates uptake of DNA but also facilitates other phenotypic changes in the pneumococcus in response to environmental changes.. Therefore, it was proposed that the competent state be renamed to the 'X-state-for-fitness' and suggested that induction of this state functions as a general response to environmental stress. Lin et al. argue that this concept is supported by the fact that *S. pneumoniae* lacks an SOS response, and that DNA damaging antibiotics trigger induction of the competent state directly. Most importantly however, Lin et al. illustrate that transcription of numerous genes involved in virulence are entirely dependent on induction of competence, or the proposed 'X-state-for-fitness'.

Therefore, while the specifics of the mechanism relating competence, autolysis and virulence are yet to be identified, there is a clear link between the three traits; induction of competence leads to lysis of a proportion of the population, and this leads to release of virulence factors into the extracellular environment which instigate disease in the infected host. While this cycle of events is supported by observation of the behaviour of the pneumococcus both *in vivo* and *in vitro*, it would seem completely counter-intuitive for evolution to favour a behaviour that is fatal to a proportion of cells within a population. What is the benefit to the pneumococcus of deliberately inducing the lysis of its own cells?

1.4.3 Overview

These are just some examples of the extensive range of virulence factors employed by *S. pneumoniae* in colonization, persistence and invasion of the human host.

The action of the autolytic enzyme in conjunction with the toxin pneumolysin presents an unusual clinical problem, in that upon treatment with antibiotics lysis of pneumococcal cells is induced, causing the symptoms of disease to temporarily worsen in response to inflammatory intracellular components that are released as a result (Hirst et al., 2004; McCullers & English, 2008; Nau & Eiffert, 2002; 2005). Working towards a better understanding of how this organism causes disease may therefore prove useful in improving treatments for those suffering from serious invasive pneumococcal disease.

This project focuses on the genetic basis for a specific aspect of pneumococcal virulence; toxicity. This phenotype was chosen because it is easily quantifiable for

GWAS, and directly contributes to virulence in this organism. Furthermore, its close relationship with autolysis and competence implies a complex network of genetic regulatory factors, which thus far is not fully understood. The work presented in this thesis contributes original knowledge to the field by using a top-down statistical approach to identify and characterise novel genetic effectors of toxicity in the two organisms *S. aureus* and *S. pneumoniae*.

1.5.0 Project aims

As part of my doctoral training programme I was required to complete two distinct rotation projects in the first year. Aim 1 below covers the work undertaken in my first rotation project, while my second rotation project focused on starting the work described in aims 2-3, which was then continued throughout the remainder of my PhD.

1) To characterise the phenotype of the toxicity-deficient *S. aureus* Δ*cyoE* mutant which was initially associated with toxicity through a genome-wide association study (GWAS).
    a. To identify the mechanism by which toxicity is lost in this mutant.
2) To apply the functional genomics approach used in previous *S. aureus* research (Laabei et al., 2014; 2015) to *S. pneumoniae*, to elucidate which genetic loci are associated with toxicity in this pathogen.
    a. To combine data from three different GWAS methods conducted by myself (with assistance from my co-supervisor) and two collaborators, such that the number of false positive and false negative results can be reduced as much as possible, and genes with the strongest associations with toxicity across more than one method can be prioritised for further investigation.
    b. To verify true-positive associations and eliminate false-positive associations by creating whole-gene deletion mutants and measuring their subsequent levels of toxicity.
3) To characterise any toxicity-deficient mutants which are identified from the previous aim with a view to identifying the mechanism by which toxicity is lost or reduced.

2. Materials and methods

2.1 List of materials used

Table 2.1. List of all materials used. All materials are detailed by name, description and supplier.

Material	Description	Supplier
0.2cm electroporation cuvette	For electroporation of competent bacterial cells	Invitrogen
1-Butanol	Solvent for butanol extraction of toxins	Fisher Scientific
6X DNA loading dye	Loading dye for gel electrophoresis	Fisher Scientific
Acetone	Solvent	Sigma
Agarose	DNA gels for electrophoresis	Acros Organics
Ammonium persulfate	Polymerising agent for SDS-PAGE	Sigma
Ampicillin	Beta-lactam antibiotic, cell wall inhibitor	Sigma
Anhydrous tetracycline	Used for pRMC2 promoter induction	Sigma
Anti-pneumolysin antibody	Polyclonal antibody to pneumolysin	Abcam (ab71810)
BHI	Brain-heart infusion broth and agar	Oxoid
Bis-polyacrylamide	40% polyacrylamide solution	Bio-Rad
Bovine serum albumin	Used as buffer in haemolytic activity assays	Sigma
Bromophenol blue	Sample buffer dye for SDS-PAGE	Sigma
β-mercaptoethanol	Used in sample buffer for SDS-PAGE gels	Sigma
$CaCl_2$	Calcium chloride	Sigma
Cell freezing media	Used for storage of THP-1 cells	Sigma
Chloramphenicol	Broad spectrum antibiotic, protein synthesis inhibitor	Sigma
CSP-2	Competence-stimulating peptide-2	Anaspec
Defibrinated horse blood	Used in haemolytic activity assay	Oxoid
DTT	Dithiothreitol	Fisher Scientific
EcoR1	Restriction enzyme	New England Biolabs

Erythromycin	Macrolide antibiotic, protein synthesis inhibitor	Sigma
Ethanol	Solvent	Fisher Scientific
FastRead counting chamber	Cell counting slides	Kova International
Fetal bovine serum	Fetal bovine serum	Gibco
Gamma-32P-ATP	Radioactive phosphorous isotope for labelling of DNA	Hartmann Analytic
Gene ruler	1kb or 50bp DNA ladder	Fisher Scientific
GeneJet Gel extraction kit	Extraction of DNA from agarose gel	Fisher Scientific
GeneJet PCR purification kit	PCR purification kit	Fisher Scientific
GeneJet Plasmid purification kit	Plasmid purification kit	Fisher Scientific
Glycerol	Used in bacterial freezer stocks and biochemical assays	Fisher Scientific
Glycine	Used in SDS-PAGE buffer	Sigma
Goat Anti-Mouse IgG H&L (HRP)	Horse-radish peroxidase coupled IgG, secondary antibody to anti-pneumolysin	Abcam (ab205719)
GoTaq 2X green master mix	PCR master mix	Promega
Guava ViaCount	Cell viability stain	Millipore
High pure PCR template extraction kit	DNA extraction kit	Roche
HisTrap Nickel binding column	Used for protein purification	GE Healthcare
HiTrap Heparin affinity column	Used for protein purification	GE Healthcare
HyBond nitrocellulose membrane	Western blot membrane	Sigma
Imidazole	Used for protein purification	Sigma
IPTG	Isopropyl β-D-1-thiogalactopyranoside	Sigma
Isopropanol	Solvent	Fisher Scientific
KCl	Potassium chloride	Sigma
KpnI	Restriction enzyme	New England Biolabs
LB	Luria-Bertani broth and agar	Fisher Scientific
L-glutamine, penicillin, streptomycin solution	Glutamine supplement and antibiotic solution for cell culture	Sigma
Lysostaphin	Glycyl-glycine endopeptidase	Cell Sciences

Lysozyme	Muramidase, used for DNA extraction	Sigma
$MgSO_4$	Magnesium chloride	Sigma
Microspin S-400	Columns used for purification of radiolabelled DNA	GE Healthcare
MonoQ anion exchange chromatography column	Used for protein purification	GE Healthcare
NaCl	Sodium chloride	Sigma
NcoI	Restriction enzyme	New England Biolabs
Nunc 96-well plate	Used in haemolytic activity assay	Fisher Scientific
Nunc 96-well plate	Used for haemolytic activity assays	Nunc
Opti4CN substrate kit	Colorimetric (HRP) substrate	Bio-Rad
Oxacillin	Beta-lactam antibiotic, cell wall inhibitor	Sigma
Page ruler plus	Pre-stained protein ladder	Fisher Scientific
PBS	Tissue-grade phosphate buffered saline	Gibco
PBS	Phosphate buffered saline	Fisher Scientific
Phusion master mix	High fidelity DNA polymerase in master mix	Fisher Scientific
PMSF	Phenylmethylsulfonyl fluoride	Fisher Scientific
Protein G-HRP	Protein G-Horseradish peroxidase conjugate	Invitrogen
Proteinase K	Broad spectrum serine protease	Sigma
qScript cDNA synthesis kit	Used for reverse transcription of RNA to cDNA	Quantabio
Qubit BR dsDNA assay kit	DNA quantification kit	Invitrogen
Quick-RNA Fungal/Bacterial Miniprep kit	RNA extraction kit	Zymo
Rabbit anti-*hla*	Polyclonal wole antiserum against α-haemolysin	Sigma
RPMI 1640	Roswell Park Memorial Institute cell culture medium	SLS/Lonza
SDS	Sodium dodecyl sulfate	Fisher Scientific
Sheep blood in Alsever's solution	Used in haemolytic activity assay	Oxoid
SimplyBlue SafeStain	SDS-PAGE gel stain	Fisher Scientific
Sodium citrate	Used to make LK agar plates	Sigma

Spectinomycin	Aminocyclitol antibiotic, protein synthesis inhibitor	Generon
Sucrose	Used in electroporation & protein purification	Sigma
SYBR-Safe	DNA gel stain	Invitrogen
T4 PNK	T4 polynucleotide kinase	New England Biolabs
T-25	Tissue culture flask	Corning
T-75	Tissue culture flask	Corning
TAE	Tris-acetate EDTA buffer	Fisher Scientific
TBE	Tris-borate EDTA buffer	Fisher Scientific
TCA	Trichloroacetic acid	Fisher Scientific
TCEP	Tris(2-carboxyethyl)phosphine	Sigma
TEMED	N, N, N, N'-tetramethylethylenediamine	Sigma
Triton X-100	Detergent	Sigma
Tris	Trisaminomethane	Sigma
Trypan blue	Viable cell stain	Sigma
Tryptone	Used to make LK agar	Sigma
TSA	Tryptic soy agar	Sigma
TSB	Tryptic soy broth	Sigma
Turbo DNA-free kit	DNAse digestion kit	Fisher Scientific
Tween-20	Detergent	Sigma
Urea	Used for resuspension of protein extractions	Sigma
Virkon	Disinfectant	Fisher Scientific
XhoI	Restriction enzyme	New England Biolabs
Yeast extract	Used to make LK agar	Fisher Scientific
Zeta-probe blotting membrane	Blotting membrane	Bio-Rad

2.2 Experiments using *Staphylococcus aureus*

Experiments were conducted by myself unless otherwise stated. Where experiments were not conducted by me, this is due to both resources and expertise being more readily available in collaborating labs regarding those individual experiments.

2.2.1 Strains and cultivation conditions used throughout

The Δ*cyoE* mutant was obtained from the Nebraska Transposon Mutant Library (mutant NE1434), which is a collection of 1,952 *S. aureus* mutants in the background strain USA300 JE2 (Bae et al., 2008; Fey et al., 2013). The USA300 JE2 wild type was used as a control in all assays. Bacterial strains were grown at 37°C in brain heart infusion broth (BHI) or on tryptic soy agar (TSA). When needed, erythromycin (5µg/ml), chloramphenicol (10µg/ml) and tetracycline (50ng/ml) were added to both plates and liquid cultures to select for and maintain the growth of mutant and complemented strains.

2.2.2 Phenotypic assays

2.2.2.1 Growth, pH and acetic acid assays

** These assays were done by Stewart Gardner, with input from Greg Somerville, at the University of Nebraska.

These assays were done to characterise the metabolic differences between the Δ*cyoE* mutant and JE2 wild type.

Bacterial strains were grown in filter-sterilized tryptic soy broth (TSB) and cultivated at 37°C, with 225 rpm aeration, and using a flask-to-medium ratio of 10:1. Bacterial pre-cultures were prepared from overnight cultures diluted 1:100 in TSB and incubated for 1.5 to 2 h. These pre-cultures were centrifuged for 5 min at 5,000 rpm, and the exponentially growing cells were inoculated into pre-warmed TSB to an optical density at 600 nm (OD_{600}) of 0.02. No antibiotics were used in this assay. Cultures were diluted prior to reading the density at the later stages of growth to avoid any saturation effects in the spectrophotometer. The pH of the culture medium was determined hourly using an Accumet AR60 pH meter (Fisher Scientific). The acetic acid assays were performed on culture supernatants (1ml) that were harvested hourly by centrifugation and acetate concentrations were determined with Acetic Acid Test Kits purchased from R-Biopharm and used according to the manufacturer's protocol.

2.2.2.2 Toxicity assay

This assay was done to characterise the toxicity phenotype of the Δ*cyoE* mutant.

THP-1 cells are an immortal monocytic cell line (Tsuchiya et al., 1980) that is sensitive to 13 of the 15 toxins produced by *S. aureus* and present in the

bacterial supernatant (Laabei et al., 2014). They are continuously sub-cultured at 2-3 day intervals in a solution of RPMI 1640 containing foetal bovine serum and an antibiotic solution of 200mM L-glutamine, 10,000 units of penicillin and 10mg/ml streptomycin. Following overnight growth in BHI broth, the bacterial cultures were centrifuged for 10 minutes at 10,000-12,000 x g and the supernatant was harvested. Supernatant was not tested to determine whether it was completely cell free in part due to the number of cultures used in each assay and the feasibility of doing so for each sample, and also as it was thought that any residual cells in the culture would not affect the outcome, as toxins are secreted by *S. aureus* into the external medium. The supernatant was diluted to a 30% vol/vol in BHI broth and 20µl of this was added to 20µl of washed THP-1 cells at a concentration of 120-150 cells per 1µl, and incubated for 12 minutes at 37°C. Following incubation of bacterial supernatant with THP-1 cells, samples were stained with 260µl Guava ViaCount reagent, incubated at room temperature for 5 minutes, and loaded onto a Guava flow cytometer to determine the percentage of THP-1 cell death in each sample.

2.2.2.4 RNAIII activity assay

This assay was done to determine the levels of activity of the the global toxicity regulator Agr in the Δ*cyoE* mutant compared to the JE2 wild type.

A plasmid containing a GFP-tagged copy of the RNAIII gene was transformed into strains USA300 JE2, Δ*cyoE*, RN6390B (an *agr*-positive strain) and RN6911 (an *agr*-negative strain). Single colonies were inoculated into BHI as described above for liquid overnight cultures, and the following morning a 1:10,000 dilution was made into BHI at a flask to medium ratio of 10:1. Strains were cultured at 37°C and aerated at 180rpm; $OD_{(600)}$ and GFP fluorescence (485/520) readings were then taken at hourly intervals over 12 hours.

2.2.2.5 Aconitase activity assay

** This assay was done by Stewart Gardner, with input from Greg Somerville, at the University of Nebraska.

This assay was done to determine the levels of activity of the TCA cyle in the Δ*cyoE* mutant compared to the JE2 wild type

Bacteria were harvested during the post-exponential growth phase (6 h) by centrifugation, suspended in ACN buffer (100M fluorocitrate, 90 mM Tris/HCl, pH

8.0), and lysed with lysing matrix B tubes and a FastPrep instrument (MP Biomedicals). The lysate was centrifuged for 5 min at 13,200 rpm at 4°C, and the aconitase activity in the cell-free lysate was measured by the method of Kennedy et al (1983). One unit of aconitase activity is defined as the amount of enzyme necessary to give a A240 min^{-1} of 0.0033.

2.2.2.6 Anti-α-toxin Western blot

This assay was done to determine whether a reduction in alpha toxin production could be observed in the Δ*cyoE* mutant, corresponding to its observed loss of toxicity in the THP-1 assay.

Proteins were precipitated from bacterial supernatant using trichloroacetic acid (TCA) at a final concentration of 20% for 1 hour on ice. Samples were then washed three times using ice cold acetone and solubilised in 100µl 8M urea. 20µl of each sample was mixed with 20µl loading dye and heated at 100°C for 2 minutes. 10µl of each sample was then subjected to 10% SDS-PAGE and separated proteins were electroblotted onto a nitrocellulose membrane using a semi-dry blotter at 15V for 30 minutes (BioRad). Membranes were blocked overnight using 3% BSA in PBS-T (containing 0.1% Tween) and were then incubated for 1 hour with rabbit polyclonal antibodies specific for α-toxin. After washing 3 times for 5 minutes with PBS, membranes were incubated for another hour with horseradish peroxidase-coupled Protein G. All incubation steps were done at room temperature. Membranes were washed twice for 20 minutes in PBS, and blots were visualised using an Opti-4CN detection kit. Band intensities were quantified using ImageJ (v 1.46r).

2.2.2.7 PSM (phenol soluble modulins) quantification

This assay was done to determine whether a reduction in PSM production could be observed in the Δ*cyoE* mutant, corresponding to its observed loss of toxicity in the THP-1 assay.

Overnight cultures were diluted 1:1000 in 50ml BHI and grown for 18 hours at 37°C with shaking (180rpm). 30ml supernatant was added to 10ml 1-butanol and these samples were incubated for 3 hours at 37°C with shaking. Samples were then centrifuged for 3 minutes and 1ml of the upper organic phase was collected. Protein samples were concentrated overnight using a SpeedVac and dried samples were then solubilised in 150µl 8M urea. Samples were loaded and run on 10% SDS-PAGE as described above and then stained using SimplyBlue

SafeStain as per the protocol. Band intensities were quantified using ImageJ (v 1.46r).

2.2.2.8 Micro-aerobic environment

This assay was done to assess whether the wild type JE2 strain grown in micro-aerobic conditions exhibited a comparable phenotype to the Δ*cyoE* mutant.

To assess the effect of microaerobic growth conditions on toxicity in the wild-type strain JE2, bacteria were cultivated using a flask-to-medium ratio of 10:8 and toxicity assays were conducted using culture supernatant as described above. All other growth conditions remained unchanged.

2.2.3 Phage transduction of transposon mutation into SH1000

This was done to illustrate the the *cyoE* gene is required for toxicity in more than one *S. aureus* strain, and its importance in toxicity is therefore not unique to JE2.

Donor cells were inoculated into liquid culture from single colonies and grown overnight, and the following day 200µl of this culture was added to 25ml BHI containing 250µl 1M $MgSO_4$ and 250µl 1M $CaCl_2$. This was grown for one hour after which 100µl phage 11 was added to the culture and grown for a further four hours minimum. Supernatant was obtained from this culture through centrifugation (12,000 x g for 3 minutes) and was then filter sterilised. Optimal plaque titre was in the range of 10^{7-10}. Next, recipient cells were grown overnight in 20ml LK broth (1% tryptone, 0.5% yeast extract, 0.7% potassium chloride), then this culture was centrifuged (2,500 x g for 10 minutes) and the pellet suspended in 1ml LK broth. To 250µl of recipient cells was added 500µl LK broth plus 10mM $CaCl_2$ and 250µl of the phage lysate from the previous step. This culture was incubated statically for 25 minutes and then with shaking at 180rpm for 15 minutes. 500µl ice cold 0.02M sodium citrate was then added, and the culture centrifuged at 10,000 x g for 10 minutes. The pellet was suspended in 500µl 0.02M sodium citrate and left on ice for 2 hours. 100µl of this was then plated neat on LKA plates containing 0.02M sodium citrate and selective antibiotic, and this was incubated for at least 20 hours at 37°C.

2.2.4 Construction of complemented mutant

This was done to demonstrate that restoration of the wild type *cyoE* gene in the Δ*cyoE* mutant rescued its toxicity, showing that it is specifically this gene that is responsible for the observed phenotype.

To confirm that loss of the *cyoE* gene was responsible for the observed loss in toxicity, the wild type *cyoE* gene was cloned into the transposon mutant. The plasmid vector pRMC2 was used because it contains a tetracycline-inducible promoter region that allows transcription of the gene of interest to be controlled. The wild type *cyoE* gene was amplified by PCR using the following primer sequences:

cyoE FW: GCTGGTACCATGAACAAATTTAAGGAG
cyoE RV: GCGAATTCAATTTCATCCTAACTTAATT

Restriction enzyme sites for KpnI and EcoR1 were added to the forward and reverse primers, respectively. The *cyoE* gene and plasmid pRMC2 were then digested with KpnI and EcoR1 and the resultant products were ligated using T4 DNA Ligase. Successfully ligated plasmids containing the wild-type *cyoE* gene were transformed into *E. coli* DH5α competent cells through electroporation, plasmid DNA was isolated and passaged through *S. aureus* RN4220, before finally being transformed into the strain JE2 Δ*cyoE* transposon mutant NE1434. For transformation through electroporation, bacteria were cultivated in BHI liquid culture to an $OD_{(550)}$ of 0.2-0.3 and washed four times in ice cold 0.5M sucrose. After the final wash bacteria were suspended in 100µl of 0.5M sucrose before being added to 20-30µl of DNA. Bacteria were incubated on ice with the DNA for 20 minutes and electroporated in 0.2cm cuvettes for 4.2-4.6 milliseconds. Following electroporation, 800µl BHI was added to the cuvettes and incubated for 1 hour at 37°C without shaking. The transformants were then plated on TSA containing 10µg/ml chloramphenicol, and for the transposon mutant strain 5µg/ml erythromycin was also added to the agar.

2.2.5 Statistical comparisons

Statistical comparisons were performed using a Student's T-test in Microsoft Excel, comparisons were two-tailed and assumed unequal variance.

2.3 Experiments using *Streptococcus pneumoniae*

2.3.1 Strains and growth conditions used throughout

Clinical strains used were all samples of the PMEN1 clone obtained from Stephen Bentley (Wellcome Trust Sanger Institute, Cambridge); most were of serotype 23F, but the collection also included samples of serotypes 6A, 15B, 19A, 19F and 23B. This clonal complex was chosen because it is a globally prevalent group of multidrug-resistant clinical isolates taken from patients

suffering from pneumococcal disease (Croucher et al., 2011). Whilst the group mainly consists of one serotype, some serotypes have undergone capsular switching, likely in response to inclusion of the 23F capsular antigen in pneumococcal vaccines. Whilst there is a high degree of genetic diversity between strains in this collection due to the high rates of recombination that take place in the pneumococcus, all are descended from the same ancestral strain (thought to have emerged around 1930). The underlying genome of each strain therefore corresponds to and can be compared to the reference strain ATCC 700669.

Strains were grown for up to 24 hours in 5% CO_2 at 37°C, on brain-heart infusion (BHI) blood agar plates containing 5% defibrinated horse blood. To obtain overnight liquid cultures several colonies were inoculated in 5ml BHI broth without blood and grown for 16-24 hours in the same incubation conditions. Some strains did not grow from single colonies; hence several colonies were used to inoculate all liquid cultures.

Control strains used for the haemolytic activity assays of the clinical strains were wild type TIGR4 (serotype 4) and a TIGR4 pneumolysin-deficient mutant. Growth conditions for both were as above, with the addition of 2-5µg/ml chloramphenicol to the media for the TIGR4 pneumolysin-deficient mutant, as this mutant was constructed with a chloramphenicol resistance cassette and as such is selected for with this antibiotic.

2.3.2 Haemolytic activity assay for clinical isolates

This was done to assess the toxicity levels of each clinical isolate for use in a subsequent GWAS.

Strains were grown overnight in liquid culture as described above. The following day strains were sub-cultured at a 1:1 ratio of overnight culture to fresh medium and grown for one hour until $OD_{(600)}$ 0.4-0.7 was reached. This represents the exponential phase of growth at which the pneumolysin toxin will be secreted into the culture medium, and optimisation assays using the wild type TIGR4 strain showed that there was no significant difference in toxicity levels within this range.

Cultures for each strain were then serially diluted 4-fold in a 96 well plate containing BSA assay buffer. This buffer consisted of 0.05g bovine serum albumin and 0.077 dithiothreitol dissolved in 50ml sterile phosphate buffered

saline. Six serial 4-fold dilutions were done for each strain, to which was added 50µl of triple washed sheep red blood cells.

The 96-well plate was incubated for one hour in 5% CO_2 at 37°C, following which it was centrifuged at 2000rpm for 10 minutes at room temperature. Supernatants from each well were transferred to a fresh 96-well plate and absorbance values read at 596nm were obtained using a FLUOstar Omega microplate reader.

2.3.3 Genome-wide association study - SNP-based method

This was done to identify statistical associations between genetic loci and the toxic phenotype.

166 isolates from the PMEN1 collection were used in a genome wide association study to identify single nucleotide polymorphisms (SNPs) which were significantly associated with toxicity. SNP data was obtained from Nick Croucher (Imperial College London), and has been published previously in work by Croucher et al. (2011). All statistical analysis was completed in the program 'R' (version 3.2.1). SNP data was first converted to a data frame of 1s and 0s, where 1s represent the presence of a SNP at any given position, relative to the reference genome ATCC 700669 (SPN23F). Missing data was converted to NA values and therefore excluded from statistical analysis. The genotype frequency threshold was set at 5% and the minor allele frequency threshold was set at 3%. Pearson's product-moment correlation was then used to identify any associations between SNP positions and high levels of haemolytic activity. P-values were then empirically derived using permutation of the dataset and converted to –log10 before being plotted onto a Manhattan plot to visualise the results.

2.3.4 Construction of *S. pneumoniae* knockout mutants

2.3.4.1 Overlap extension PCR

** This protocol was adapted from the methods published by Ge & Xu (2012) and Chen et al. (2015).

This was done to create whole gene deletions of genes in which specific loci had been statistically significantly associated with toxicity in a GWAS, for the purposes of verifying genes which when disrupted cause a loss in toxicity.

DNA from the reference strain ATCC 700669 (SPN23F) was extracted using a High Pure PCR Template Preparation Kit. The plasmid pVA838, containing the

erythromycin resistance cassette *ermAM*, was extracted from *E. coli* DH5α using a GeneJet Plasmid Extraction kit, and the *ermAM* cassette amplified through PCR using the following primers:

ermAM FW: GTAATTAAGAAGGAGTGATTACATG
ermAM RV: GAATTATTTCCTCCCGTTAAATAATAG

For each knockout mutant, six primers were designed to amplify three fragments in separate PCR reactions, which were then combined in a further PCR reaction to construct a single linear insert. This insert was then used to transform the reference strain through homologous recombination.

F1 and R1 primers amplified the 1kb region upstream of the gene to be knocked out, F2 and R2 primers amplified the *ermAM* cassette, and F3 and R3 primers amplified the 1kb region downstream of the gene. The F2 and R2 primers were designed to overlap with R1 and F3, respectively, such that the three fragments would stitch together to create a linear construct (see figure 2.1). All primers were designed to have an annealing temperature as close to 60°C as possible to aid the second PCR step in which the three fragments were stitched together. To reduce any potential downstream effects on other genes, for example to protect potential promotor regions and ribosome binding sites used by the neighbouring gene, the first 6 bases after the start codon and the last 30 bases before the stop codon were retained. Phusion high fidelity polymerase was used to amplify each fragment and the final linear construct.

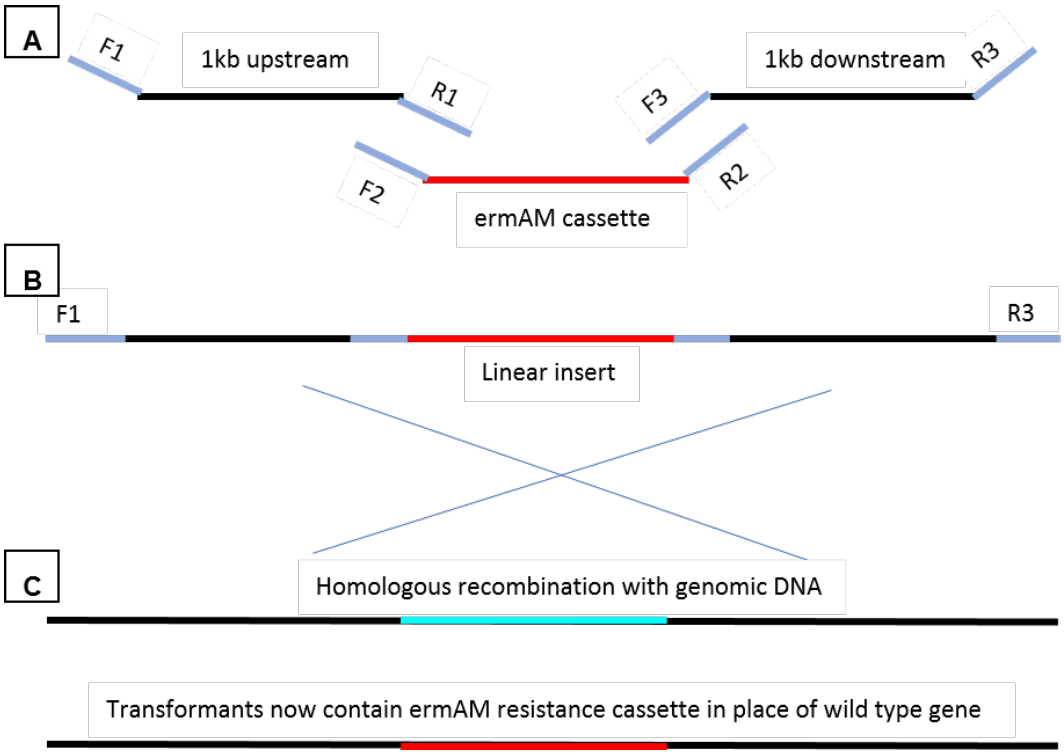


Figure 2.1: Overlap extension PCR. Method used to create linear insert, which is then transformed into reference strain through homologous recombination. **A.** Amplification of three individual fragments with overlapping R1/F2 and R2/F3 primers. **B.** PCR to stitch together three fragments to construct a linear insert, which is then used to create a whole-gene deletion of the target gene (**C**).

2.3.4.2 Transformation of linear insert into ATCC 700669 SPN23F

** This protocol has been adapted from a method used by Jane Brittan and Angela Nobbs at the Oral and Dental hospital, Bristol.

Bacteria were inoculated into 5ml BHI broth, from which 10-fold serial dilutions were made into three further tubes. These were incubated overnight. The tube containing the highest dilution of cells that had successfully grown (therefore the lowest concentration of cells) was used to inoculate 2ml of growing culture into a fresh pre-warmed tube of 20ml BHI broth. This was incubated for 2 hours. 0.5ml of this culture was added to 9.5ml pre-warmed BHI broth and incubated for 30 minutes. 1ml of this culture was transferred to at least two sterile Eppendorf tubes, to which 10µl of 10µg/ml competence-stimulating peptide-2 (CSP-2) was added each. These were incubated for 10-15 minutes, following which 200µl portions were dispensed into fresh Eppendorf tubes containing either 2µl of DNA or sterile water for the negative controls. These were incubated for a further 2 hours. 100µl of these cultures were dispensed into petri dishes and mixed with

10ml molten BHI agar containing 5% defibrinated horse blood. Two plates were used for the negative control; one to which antibiotics were added in a later step, and one to which antibiotics were not added. This is to test firstly that the antibiotic does not permit growth of non-transformed cells, while also demonstrating that cells are viable. Once set the plates were incubated for 2 hours. Plates were then overlaid with a further 10ml molten agar containing 5% defibrinated horse blood, to which erythromycin (2µg/ml) was added, except for the second negative control. Plates were incubated for up to five days at 37°C, or until colonies appeared within the agar.

For the ΔSPN23F12470(*aad9*) mutant constructed with the spectinomycin cassette *aad9*, *aad9* was amplified from the plasmid pFW5 using the following primers:

FW: GGAGGATATATTTGAATACATAC
RV: TTATAATTTTTTTAATCTGTTATTTAAATAG

Spectinomycin was added to transformation plates for this mutant at a final concentration of 200µg/ml to select for successfully transformed mutants.

2.3.4.3 Primer design for insert construction using the ply *gene as an example*

The following sequence contains the 1kb sequence upstream of the *ply* gene, the first nine bases of the *ply* gene (highlighted in yellow), the sequence encoding the erythromycin resistance cassette *ermAM* (highlighted in grey), the last 33 bases of the *ply* gene (highlighted in yellow), and the 1kb sequence downstream of the *ply* gene. Primer sequences are highlighted in red.

5’
CGCCCTTGCTCTGGTTAAAAAAAGAAGCCGATAAGGAAAAGATGAGCGCGACCCAG
TGCCAGATATTTTTCTTTTACTATCAAATGTTTGAGCTCTTATTTGCTAGAGAAAGCGA
CATGAAAGACTTATGTCTGGGAACGAAAGGTTTTTATTTCTCGCAGTTAGAGAAAAAT
TTGCTTTCTGGAGTTTCCCGATTTCTAAAAAACTTGGAGGGGAAAGTAACTCTCAAG
GCTAACCAAGAAGTATCAGCTCGCAAAGCCCTTTTTCTAGCCTTGACAACTAGCCAA
TCAGATTGGCAGGAGTTAGCTCCTGTTTTTGATTTTTATCAGACTATCGGGAGGCTT
GAAAATCCTTCTCTCTTGAGTTCTCAGGACAGACAACATCTGATGTGGATTTACCAGT
CAGCTTTGGAGAAGGATTATATTGTCAAGGTTATTGGCGACAAGCATTTTGTATTGAA
GAGACAAGATGCTACTAAATTGACAGCGCGCCAAACTCAAACTTTGGAAATTCTGAG
TCAATCAGAAGACTTGGTCAATCCTGTCTATGTTACATTAGGAGAAAAGGGGGTGCT
CTTGCTTGATTAAGAGAGGAGATGTTGTAGTTCTTTATTTGCCTTTTCCGACTATTAG
TAGCGATTTGGCTGTGAAGAATCATATGTATATCTGTATTGACAACAGCATGACTAAA
AACAAAGAGTTGGTTAAAAATCAGACCTTCAAACCAGCTCTCTTGACCAGACGTTTG
GTCAAGAACTTTATGATAGAAGAGCCGGATCTAGCTCGTAATCCTTTTACAAGACCA
ACCTTGATTGACTTAGATAAGGTATTTATGTTGGATAATACGGTTATTCCGACTTCTTA
TCTAGCCAGACGGCGACGCAATGTCTCAGAAGAATTGTACGAGGAAATTTTGGATCA
CTTAGTCCAACCACGGCTGATTTCGCTGAACAAGTCTGAGTTTATGCAACTCAATCC

AGGAACTTATTAGGAGGTAGAAGATGGCAAATATGAACAAAAATATAAAATATTCT
CAAAACTTTTTAACGAGTGAAAAAGTACTCAACCAAATAATAAAACAATTGAATTTAAA
AGAAACCGATACCGTTTACGAAATTGGAACAGGTAAAGGGCATTTAACGACGAAACT
GGCTAAAATAAGTAAACAGGTAACGTCTATTGAATTAGACAGTCATCTATTCAACTTA
TCGTCAGAAAAATTAAAACTGAATACTCGTGTCACTTTAATTCACCAAGATATTCTACA
GTTTCAATTCCCTAACAAACAGAGGTATAAAATTGTTGGGAGTATTCCTTACCATTTA
AGCACACAAATTATTAAAAAAGTGGTTTTTGAAAGCCATGCGTCTGACATCTATCTGA
TTGTTGAAGAAGGATTCTACAAGCGTACCTTGGATATTCACCGAACACTAGGGTTGC
TCTTGCACACTCAAGTCTCGATTCAGCAATTGCTTAAGCTGCCAGCGGAATGCTTTC
ATCCTAAACCAAAAGTAAACAGTGTCTTAATAAAACTTACCCGCCATACCACAGATGT
TCCAGATAAATATTGGAAGCTATATACGTACTTTGTTTCAAAATGGGTCAATCGAGAA
TATCGTCAACTGTTTACTAAAAATCAGTTTCATCAAGCAATGAAACACGCCAAAGTAA
ACAATTTAAGTACCGTTACTTATGAGCAAGTATTGTCTATTTTTAATAGTTATCTATTAT
TTAACGGGAGGAAATAACCTCAGGTAGAAGATAAGGTAGAAAATGACTAGGAGAG
GAGAATGCTTGCGACAAAAAGAGGCGATGATCTCTCTGCGGATATCTGAAAGTTTAT
CTCTTGCCTAGCGATTTCCATTGGAGTATGCTCGTGATGTTAAAATGAATCTATCAAA
CAGAATGTGTAAAAAAGTAAAAAGGGTAGCCAATTCCTTGATATACAGGGGATTGAT
TATCCTTTACTTGTTTTAGATGATTGGAAATTCAAGAATGACACGTATCGGAAATATTT
TGCTGTTTATATCTAATACCCATTTATGGAAAAAATGAAATTAAATATGAACAAATCAA
TTGCTAGCAATGGTTTAGAAGTGCCAGTGTACTAGTCTAATTTCTCCATCTTGGGAAC
ACGTAAATTATACTCTTCGAAAATCTCTTCAAACCAGGTCAGCCCTATCCGCAACCTC
AAAACAGTGTTTTGAGCAACCTGCGGCTAGCTTTCTAGTTTGCTTTTTGATTTTTATT
GAGTATTAGAGGTTTGCTTAGAAACTCTTCTATTGATTTCTTGGAATACAATAGCGCC
TATAAGAACTGTTGGATTTCAAAGTATTTACTTGAACTATGATATGTCTGGAGTAGAC
TGTCAACAAAAGGTGCAGTATGATTGTTTTTGTCGCTTATATCTACAAACTTCAAACA
GTAATTTGAGCTGACTTTGTCAAACTTGGCTGAAACTTTCTATTTGCAGATAATGGTC
TAGTTTGTTTTTTGAATGTGGATGAGTGTCTGTTGATAAAAAGATAGAAAGGCAAGAC
CATCTTAGCAGGTAAAGTGCTATCCTAGCTTGATTTGCTAATAAAATCCATTTCCTCT
AGTGAAAATCAAAAAAACTTGTGTTATAATAAGAAAGATTAAAATGTGAAAAAAGGAG
ATTCCTAATGGGACGTAAATGGGCCAATATCGTAGCCAAGAAAACGGCTAAAGATGG
AGCCAACTCTAAAGTATATGCAAAATTTGGTGTAGAAATCTATGTAGCAGCTAAAAAA
GGTGATCCAG

3’

Overlapping primers were designed to amplify three individual fragments which could then be stitched together to form a single linear insert containing the *ermAM* cassette in place of the wild type gene of interest, as described in section 2.3.4.1 above. Example primers for the *ply* gene are as follows:

Fragment 1 primers:

F1: 5’ CGCCCTTGCTCTGGTTAAAAAAAGAAGC 3’

R1: 5’ GGAGGTA**GAAG**ATGGCAAATATGAACAAAAATAT 3’

Reverse complement for R1 primer:

5’ ATATTTTTGTTCATATTTGCCATCTTCTACCTCC 3’

Fragment 2 primers:

F2: 5’ GAAGATGGCAAATATGAACAAAAATATAAAATATTC 3’

R2: 5’ AACGGGAGGAAATAACCTCAGGTAGAAG 3’

Reverse complement for R2 primer:

5’ CTTCTACCTGAGGTTATTTCCTCCCGTT 3’

Fragment 3 primers:

F3: 5' **GGAGGAAATAACCTCAGGTAGAAG**ATAAG 3'
R3: 5' CTATGTAGCAGCTAAAAAAGGTGATCCAG 3'
Reverse complement for R3 primer:
5' CTGGATCACCTTTTTTAGCTGCTACATAG 3'

Following amplification of each individual fragment, all three were stitched together in one PCR reaction using the F1 and R3 primers. Each insert was then transformed into *S. pneumoniae* strain ATCC 700669 (SPN23F) using the method described in section 2.3.4.2. Successful mutants were verified using PCR checks as described in section 2.3.4.5 below.

2.3.4.5 Primer design for verification of successfully transformed inserts using the ply *gene as an example*

For verification of successfully transformed mutants, primers were designed for three different PCR checks:

1) Confirmation that the mutant is not a contaminating organism using primers to the wild type SPN23F *ply* gene (except for the Δ*ply* mutant, in which primers to the wild type SPN23F13160 gene were used instead)
2) To verify the absence of the wild type gene that has been replaced with the antibiotic resistance cassette
3) To verify uptake of the 3kb insert in the correct position, using primers for the region 1kb upstream of the 3kb insert to within the antibiotic resistance cassette

1) Primers to the wild type *ply* gene used for the first PCR check for all mutants except Δ*ply* were:

FW: GCAGTAAATGACTTTATACTAGCTATGATAAAC
RV: GAGAGTTGTTCCCCAAATAGAAATCG

Primers used to verify the Δ*ply* mutant only for this step were to the wild type SPN23F13160 gene as follows:

FW: ATGAAACGAATTACCGCAAATCAATATC
RV: AACCACCTCCTTTTCTAACACGATTATTC

2) Primers for the second PCR check were designed as follows, with the forward primer starting at least 6bp after the start codon and the reverse primer starting at least 30bp before the stop codon, since these bases were not deleted in the mutants.

Wild type *ply* gene sequence:

5'
ATGGCAAATAAAGCAGTAAATGACTTTATACTAGCTATGAATTACGATAAAAAGAAAC
TCTTGACCCATCAGGGAGAAAGTATTGAAAATCGTTTCATCAAAGAGGGTAATCAGC
TACCCGATGAGTTTGTTGTTATCGAAAGAAAGAAGCGGAGCTTGTCGACAAATACAA
GTGATATTTCTGTAACAGCTACCAACGACAGTCGCCTCTATCCTGGAGCACTTCTCG
TAGTGGATGAGACCTTGTTAGAGAATAATCCCACTCTTCTTGCGGTCGATCGTGCTC
CGATGACTTATAGTATTGATTTGCCTGGTTTGGCAAGTAGCGATAGCTTTCTCCAAGT
GGAAGACCCCAGCAATTCAAGTGTTCGCGGAGCGGTAAACGATTTGTTGGCTAAGT
GGCATCAAGATTATGGTCAGGTCAATAATGTCCCAGCTAGAATGCAGTATGAAAAAA
TCACGGCTCACAGCATGGAACAACTCAAGGTCAAGTTTGGTTCTGACTTTGAAAAGA
CAGGGAATTCTCTTGATATTGATTTTAACTCTGTCCATTCAGGCGAAAAGCAGATTCA
GATTGTTAATTTTAAGCAGATTTATTATACAGTCAGCGTAGATGCTGTTAAAAATCCA
GGAGATGTGTTTCAAGATACTGTAACGGTAGAGGATTTAAAACAGAGAGGAATTTCT
GCAGAGCGTCCTTTGGTCTATATTTCGAGTGTTGCTTATGGGCGCCAAGTCTATCTC
AAGTTGGAAACCACGAGTAAGAGTGATGAAGTAGAGGCTGCTTTTGAAGCTTTGATA
AAAGGAGTCAAGGTAGCTCCTCAGACAGAGTGGAAGCAGATTTTGGACAATACAGA
AGTGAAGGCGGTTATTTTAGGGGGCGACCCAAGTTCGGGTGCCCGAGTTGTAACAG
GCAAGGTGGACATGGTAGAGGACTTGATTCAAGAAGGCAGTCGCTTTACAGCAGAT
CATCCAGGCTTGCCGATTTCCTATACAACTTCTTTTTTACGTGACAATGTAGTTGCGA
CCTTTCAAAATAGTACAGACTATGTTGAGACTAAGGTTACAGCTTACAGAAACGGAG
ATTTACTGCTGGATCATAGTGGTGCCTATGTTGCCCAATATTATATTACTTGGGATGA
ATTATCCTATGATCATCAAGGTAAGGAAGTCTTGACTCCTAAGGCTTGGGACAGAAA
TGGGCAGGATTTGACGGCTCACTTTACCACTAGTATTCCTTTAAAAGGGAATGTTCG
TAATCTCTCTGTCAAAATTAGAGAGTGTACCGGGCTTGCCTGGGAATGGTGGCGTAC
GGTTTATGAAAAAACCGATTTGCCACTAGTGCGTAAGCGGACGATTTCTATTTGGGG
AACAACTCTCTATCCTCAGGTAGAAGATAAGGTAGAAAATGACTAG
3'

Primers to the wild type *ply* gene (highlighted in red in sequence above):

FW: GCAGTAAATGACTTTATACTAGCTATGATAAAC
RV: GAGAGTTGTTCCCCAAATAGAAATCG

3) Primers for the third PCR check were designed as follows.

Region 1kb upstream of insert to within *ermAM* cassette, which is highlighted in grey:

5'
CTAATTGTTTTTTTAGGGCGATTGATTTGTACTTCCGTCGCATCAATCATTACCGTGT
CCTCAGAACTGAGAGGAGTTCTTGAAATCGTAACACCACTTTGAACAAGAGTTACTT
CAACCCATTGGCTCCGACGGATTAAGTTGCTTTCGTGAATACCAAAATCAGCCGCAA
TTTGTTCATAAGTTCGATATTCTCGCACATATTGAAGAGTGGCCATAAGAAGGTCTTC
TAGGTTTAATTTAGGTTTTCGTCCACCTTTTGCGTGTTTAAGTTGATAAGCTGTTTTTA
ATACAGCTAACATCTCTTCAAAAGTGGTACGCTGAACACTAACAAGACGCTTAAATTG
TGCATCAGTTAGTTGTTTACTTGCTTCATAATTCATAGAACTACTATACCATATTTTGT

TTCGTAGGAAGTCTAATAAGTTATAGTCCAGTTTTTACTATACTTTGATATTCAAATAT
GTTATACTATGATTATCAAAGTAAATGAAAGGAGTTACGAATATGACGCCAGAAGAAA
TGTACCTGACAGAGCGATTAGACGTACAGATAGCTCATTTTTTAAAGAAAAGCGTTCA
ACATCGTAGGCGCTATAAGGTATTAAAAATAACAGAAATCGTGGCAGGTTTCCTCAT
AGCTGTCTTTTGTGCTATTCCTATGCCAGGTGATCGCTACCGTTTGATTTCGGTTGC
CTTATCCAGTCTCGGCTTGCTGTGTGAGGGGATTATCAATTTGTATAATGCAAAGGA
AAATTGGATTTCTTACCAAAAAACTGCGCAACTCCTGGAGAAAGAAAAATTCCTCTAT
CAATGCCAAACGGAGAAATATGCAGGAAAGACCAAGGCTTTTGCCCTATTTGTCAAG
ACATGCGAAGGTCTTATCTCAGAGGAGATTAACCAGTGGGAAAGTATCCAGTCAAAA
GAAGTGGCAGCTAGTGCAGATGCTCCAGTTAAAAAAGAGTAGGAGGTAGAGGAAAT
GTCTCAATCCAGCTACCTGTCGCCCTTGCTCTGGTTAAAAAAAGAAGCCGATAAGGA
AAAGATGAGCGCGACCCAGTGCCAGATATTTTTCTTTTACTATCAAATGTTTGAGCTC
TTATTTGCTAGAGAAAGCGACATGAAAGACTTATGTCTGGGAACGAAAGGTTTTTATT
TCTCGCAGTTAGAGAAAAATTTGCTTTCTGGAGTTTCCCGATTTCTAAAAAACTTGGA
GGGGAAAGTAACTCTCAAGGCTAACCAAGAAGTATCAGCTCGCAAAGCCCTTTTTCT
AGCCTTGACAACTAGCCAATCAGATTGGCAGGAGTTAGCTCCTGTTTTTGATTTTTAT
CAGACTATCGGGAGGCTTGAAAATCCTTCTCTCTTGAGTTCTCAGGACAGACAACAT
CTGATGTGGATTTACCAGTCAGCTTTGGAGAAGGATTATATTGTCAAGGTTATTGGC
GACAAGCATTTTGTATTGAAGAGACAAGATGCTACTAAATTGACAGCGCGCCAAACT
CAAACTTTGGAAATTCTGAGTCAATCAGAAGACTTGGTCAATCCTGTCTATGTTACAT
TAGGAGAAAAGGGGGTGCTCTTGCTTGATTAAGAGAGGAGATGTTGTAGTTCTTTAT
TTGCCTTTTCCGACTATTAGTAGCGATTTGGCTGTGAAGAATCATATGTATATCTGTA
TTGACAACAGCATGACTAAAAACAAAGAGTTGGTTAAAAATCAGACCTTCAAACCAG
CTCTCTTGACCAGACGTTTGGTCAAGAACTTTATGATAGAAGAGCCGGATCTAGCTC
GTAATCCTTTTACAAGACCAACCTTGATTGACTTAGATAAGGTATTTATGTTGGATAAT
ACGGTTATTCCGACTTCTTATCTAGCCAGACGGCGACGCAATGTCTCAGAAGAATTG
TACGAGGAAATTTTGGATCACTTAGTCCAACCACGGCTGATTTCGCTGAACAAGTCT
GAGTTTATGCAACTCAATCCAGGAACTTATTAGGAGGTAGAAGATGGCAAATATGAA
CAAAAATATAAAATATTCTCAAAACTTTTTAACGAGTGAAAAAGTACTCAACCAAATAA
TAAAACAATTGAATTTAAAAGAAACCGATACCGTTTACGAAATTGGAACAGGTAAAGG
GCATTTAACGACGAAACTGGCTAAAATAAGTAAACAGGTAACGTCTATTGAATTAGAC
AGTCATCTATTCAACTTATCGTCAGAAAAATTAAAACTGAATACTCGTGTCACTTTAAT
TCACCAAGATATTCTACAGTTTCAATTCCCTAACAAACAGAGGTATAAAATTGTTGGG
AGTATTCCTTACCATTTAAGCACACAAATTATTAAAAAAGTGGTTTTTGAAAGCCATG
CGTCTGACATCTATCTGATTGTTGAAGAAGGATTCTACAAGCGTACCTTGGATATTCA
CCGAACACTAGGGTTGCTCTTGCACACTCAAGTCTCGATTCAGCAATTGCTTAAGCT
GCCAGCGGAATGCTTTCATCCTAAACCAAAAGTAAACAGTGTCTTAATAAAACTTACC
CGCCATACCACAGATGTTCCAGATAAATATTGGAAGCTATATACGTACTTTGTTTCAA
AATGGGTCAATCGAGAATATCGTCAACTGTTTACTAAAAATCAGTTTCATCAAGCAAT
GAAACACGCCAAAGTAAACAATTTAAGTACCGTTACTTATGAGCAAGTATTGTCTATT
TTTAATAGTTATCTATTATTTAACGGGAGGAAATAACCTCAGGTAGAAGATAAGGTAG
AAAATGACTAG

3’

Primers to verify presence of insert at correct location within genome (highlighted in red in sequence above):

FW: ATCTCTTCAAAAGTGGTACGCTGAACACTAAC
RV: GTGTTCGGTGAATATCCAAGGTACGCTTG

2.3.4.4 Primers used to construct mutants

A full list of the primers used to construct whole-gene deletion mutants, and the primers used to verify successfully transformed mutants, is provided in appendix I.

2.3.5 Characterisation of toxicity-deficient mutants

2.3.5.1 Haemolytic activity assay for ATCC 700669 (SPN23F) whole gene deletion mutants

Haemolytic activity assays were conducted as described above with a few minor changes and additions to characterise the toxicity phenotype of successfully constructed mutants in the ATCC 700669 SPN23F background.

Overnight cultures were added to the 96-well plate containing BSA assay buffer without sub-culturing for 1 hour. This was done because it was established while assaying the clinical strains that sub-culturing made little difference to the outcome of the assay, as toxins which were present in the supernatant of the overnight culture will be carried over to the sub-culture the following morning. The clinical strains were only sub-cultured because it was initially thought that cells would need to be grown to early exponential phase on the day of the assay for the toxicity to be observed. It was also established through several experiments with the wild type strain that the same level of toxicity was observed within the range of 0.2-0.9 $OD_{(600)}$. Since this strain grows only to a maximum $OD_{(600)}$ of around 1, it was established that past the point of $OD_{(600)}$ 0.2, there was no significant difference in toxicity, making the 1 hour sub-culture unnecessary. Haemolytic activity was measured by reading absorbance at 490nm.

Haemolytic activity of the supernatant was also assayed using the mutant strains, primarily for comparison to the whole cell culture method and also to specifically investigate whether pneumolysin is being released from cells into the external medium. It was not assayed for the clinical strains as the whole cell culture method was deemed to be sufficient for observing variations in toxicity and it would have been too time consuming to measure both supernatant and whole cell culture for each strain. Supernatant was taken from overnight cultures by centrifugation at 3000 x *g* for 10 minutes. Six serial 2-fold dilutions were then made in BSA assay buffer in a 96-well plate, with the first well remaining undiluted, and the haemolytic activity assay was conducted as described above with no further changes.

2.3.5.2 Contact-dependent haemolytic activity assay

This assay was done to characterise (by proxy) the activity of pneumolysin toxin that is specifically localised to the bacterial cell wall.

All bacterial strains which were tested for contact-dependent haemolysis, including whole gene deletion mutants and the ATCC 700669 (SPN23F) wild type, were grown overnight in 10ml and harvested by centrifugation. All centrifugations were done at 3000 x *g* for 7 minutes at 4°C unless stated otherwise. Supernatant was discarded, cells were washed in 2ml ice-cold PBS and harvested by centrifugation. Strains were resuspended in ice-cold PBS to $OD_{(600)}$ 1. The maximum volume which could be taken equivalently from all strains was harvested by centrifugation, and cell pellets were resuspended in 400µl ice-cold PBS. Red blood cells were harvested by centrifugation from 1ml fresh sheep blood. Supernatant was discarded and erythrocytes were resuspended in 3ml PBS to give a final concentration of 25% cells/volume. 50µl bacterial suspension was mixed with 50µl 25% erythrocytes and centrifuged at 9600 x *g* for 5 minutes at 4°C to achieve close contact between bacterial cells and erythrocytes. Controls used were: bacterial and erythrocyte suspensions not centrifuged; bacteria centrifuged while blood cells not centrifuged and vice versa; bacteria replaced by PBS only and blood replaced by PBS only. Samples were then incubated statically for 3 hours at 37°C. After incubation, 70µl of supernatant was discarded and pellets resuspended in 150µl ice cold PBS. Samples were centrifuged at 9600 x *g* for 1 minute at 4°C, and 100µl aliquots of supernatant were transferred from each sample to a 96-well plate. Haemolytic activity was measured by reading absorbance at 490nm.

2.3.5.3 Anti-pneumolysin Western blot

** Protein extraction protocols for supernatant and cell pellets were taken from Balachandran et al. (2001) as follows:

This was done to determine whether there were differences in the amount of pneumolysin produced in wild type and mutant strains.

Proteins were precipitated from the bacterial supernatant of 15ml overnight cultures using trichloroacetic acid (TCA) at a final concentration of 10%, incubated overnight at 4°C. Samples were centrifuged at 3000 x *g* for 15 minutes and washed twice using ice cold acetone. Samples were then resuspended in 400µl sterile water and stored at -20°C. Proteins were extracted from the cell pellets of 15ml overnight cultures using a solution of 0.01% SDS, 0.1% sodium deoxycholate and 0.015M sodium citrate, which was added at a volume corresponding to 10% of the original culture volume (1.5ml). Samples were incubated in this solution for 30 minutes at 37°C and then stored at -20°C. SDS-

PAGE gels and Western blotting were conducted as described in section 2.2.2.6 above, with the exception that anti-pneumolysin was used as the primary antibody, and goat anti-mouse IgG H&L (HRP) was used as the secondary antibody (see list of materials).

2.3.5.4 Autolysis assay

This was done to determine whether differences in autolysis were responsible for observed differences in toxicity between wild type and mutant strains.

5ml liquid cultures were inoculated and grown for 18-20 hours. Cultures were then washed once in ice cold PBS and resuspended in 5ml ice cold PBS. Optical density readings were taken for each culture at this point at $OD_{(600)}$. Triton X-100 was added to a final concentration of 0.1% for each culture following this OD_{600} reading and cultures were returned to the incubator. Every 15 minutes OD_{600} readings were taken for each culture up to the 1 hour 15 minute time point, at which point OD_{600} readings for the wild type strain began to plateau. Autolysis-deficient mutants retain high optical density throughout the assay in comparison.

*2.3.5.5 Comparison of amount of ICE**Sp**23FST81 element present in circularised form between ΔSPN23F12470 mutant and isogenic wild type*

This was done to determine whether mobilisation of the ICE*Sp*23FST81 was affected in the ΔSPN23F12470 mutant.

Plasmids were extracted from 5ml overnight cultures for ΔSPN23F12470 and the ATCC 700669 (SPN23F) isogenic wild type strain using a GeneJET Plasmid Purification kit. Primers were designed firstly to amplify the genes at the start and end of the ICE*Sp*23FST81 element, as defined by Croucher et al. (2009), using the following primers:

SPN23F12410 FW: CAAAGACAAAGAAATATAAAGGTGTATAC
SPN23F12410 RV: TCAACTCTGGCAGGAATTTGGTAG

SPN23F13160 FW: ATGAAACGAATTACCGCAAATCAATATC
SPN23F13160 RV: AACCACCTCCTTTTCTAACACGATTATTC

Following successful amplification of these genes to demonstrate that the primers worked, the SPN23F12410 FW and SPN23F13160 RV primers were used to amplify the junction of the circularised ICE element where the two linear ends join in both the wild type strain and the ΔSPN23F12470 mutant, to confirm it was

present in its circularised form in both strains. The amount of circularised ICE element present in the ΔSPN23F12470 mutant was then compared to its isogenic wild type by PCR using three separate plasmid extractions from three independent cultures for each strain. Plasmid extractions for each strain were standardised to a quantity of 20ng/µl, of which 1µl was used as template for each PCR.

2.3.5.6 qRT-PCR

This was done to investigate differences in transcript expression of key virulence-related genes in the ΔSPN23F12470 mutant compared to its isogenic wild type.

Two separate cultures were grown per strain for 18 hours to stationary phase in 5ml BHI with antibiotics where required. RNA was extracted using a Zymo Quick-RNA Fungal/Bacterial Miniprep kit following the manufacturer's protocol. Excess DNA was digested using a Turbo DNase digest kit, as per the manufacturer's protocol. RNA was then quantified and its quality determined using a Nanodrop. RNA quality was determined from the A260/280 ratio. RNA was reverse transcribed to cDNA using a qScript cDNA synthesis kit, as per the manufacturer's protocol. RNA samples were standardised such that 100ng/µl was added to each cDNA synthesis reaction. cDNA was subsequently quantified using a Nanodrop and where necessary was standardised again to ensure the same quantity in each sample. qPCR was performed using a Mic qPCR cycler (Bio Molecular Systems) and reactions were set up using a SensiFAST SYBR No-ROX master mix. Three technical repeats were conducted for each cDNA sample. The following primer pairs were used for each strain, with *recA* used as the housekeeping gene:

ply FW: GGCTGATTTCGCTGAACAAG
ply RV: CCTTGAGTTGTTCCATGCTG

lytA FW: TTATCACTGGCGGAAAGACC
lytA RV: TCAGTTCAACCGCTGCATAG

comA FW: TCACCGCTTGACTATTGCTG
comA RV: GCATGCTTTCCTTCTTCGAC

comE (SPN23F22680) FW: CAAGACAACGGGAAAAGTCC
comE (SPN23F22680) RV: AGCTGAGCCACTTCAAATCC

recA FW: ATCGGAGATAGCCATGTTGG
recA RV: ATAGAGGCGCCAAGTTTACG

qRT-PCR data were analysed using the $2^{-\Delta\Delta CT}$ method. Negative controls of water-only and each original RNA sample were used to ensure there was no contamination of either cDNA or the reaction mix.

2.3.5.7 Complementation of ΔSPN23F12470 mutant

This was attempted to demonstrate that restoration of the wild type SPN23F12470 gene in the ΔSPN23F12470 mutant rescued its toxic phenotype. Unfortunately, attempts to complement this mutant have thus far been unsuccessful.

Complementation of the ΔSPN23F12470 deletion was attempted using the vector pMSP7517 (provided by Angela Nobbs, University of Bristol), as it contains a nisin-inducible promoter to allow transcription of the complemented gene to be controlled. pMSP7517 contains a gene (*prgB*) flanked by restriction sites for NcoI and XhoI only, however, XhoI digests SPN23F12470 at a site 1,857 bases into the gene. In the absence of an alternative plasmid, site directed mutagenesis was conducted to change the digest site sequence from CTCGAG to CTGGAG, which is a synonymous change that does not cause a change in the resulting amino acid. Site directed mutagenesis was done by PCR amplification of the wild type SPN23F12470 gene in two separate overlapping fragments, using the following primer sequences:

F1: CCATGGAGGAGGATGATACATGATGACAACATTAGATTTTAAAAC
R1: CAAACTCCAGTCCTTTGGAAC
F2: CAAAGGACTGGAGTTTGATC
R2: CTCGAGTTAATTAACTAAATTACTTTCTTTTC

Primers F1 and R2 were then used to stitch together the two fragments, resulting in a wild type SPN23F12470 gene with a single, synonymous base change to eliminate the XhoI digest site in the middle of the gene (designated SPN23F12470a). Restriction enzyme sites for NcoI and XhoI were added to the F1 and R2 primers, respectively. The SPN23F12470a gene and pMSP7517 vector were subsequently digested with NcoI and XhoI and the resulting products were gel extracted and ligated using T4 DNA ligase. Ligations were transformed into *E. coli* DH5α using heat shock: 5µl ligation mix was added to 50µl competent cells and incubated for 30 minutes on ice, cells were then incubated at 42°C for two minutes followed by a further 5 minutes on ice, after which 750µl BHI broth was added and cells were incubated for 45 minutes at 37°C with shaking. Cells

were then centrifuged for 1 minute at 10,000 x g, and pellets resuspended in 100µl supernatant for plating onto BHI containing 200µg/ml erythromycin.

No successful transformants have been obtained following transformation into *E. coli* DH5α. Transformation of the ligation mix was also attempted directly into *S. pneumoniae* ATCC 700669 ΔSPN23F12470(*aad9*) as per the protocol described in section 2.3.4.2 above, with no success.

2.3.5.8 Purification of SPN23F12470 wild type protein

** This was done on my behalf by Oliver Wilkinson, Roz Williamson and Elliott Mainstone in the Dillingham lab.

This was done to isolate wild type SPN23F12470 protein for use in a subsequent electrophoretic mobility shift assay.

SPN23F12470 protein was purified from BL21(DE3) *E. coli* cells transformed with the plasmid pET15b containing the SPN23F12470 gene modified with a 6x Histidine tag at its N-terminus. Cells were grown at 37°C in LB broth containing 100µg/ml ampicillin to an $OD_{(600)}$ of around 0.5, whereupon cells were temperature acclimatised to 18°C and gene expression was induced for 12 hours via the addition of 1mM IPTG. Cells were harvested by centrifugation for 20 minutes at 6000rpm and resuspended in buffer containing 50mM Tris pH 7.5, 150mM NaCl, 10% sucrose and 0.1mM PMSF. Cells were lysed by sonication in buffer containing 1mM TCEP, 500mM NaCl, 20mM imidazole and a protease inhibitor cocktail, and cell debris was removed by centrifugation for 30 minutes at 21,500rpm. Cell lysate was loaded onto a 5ml HisTrap Nickel binding column which had been equilibrated in HisB buffer (20mM Tris pH7.5, 500mM NaCl, 1mM TCEP, 5% glycerol) plus 20mM imidazole. Proteins possessing His-tags were eluted with a gradient from 20mM to 500mM imidazole over 32 minutes with a flow rate of 3ml/min. Protein elution was monitored by measuring absorbance at 280nm. Fractions containing the most protein were pooled and applied to a HiTrap Heparin affinity column equilibrated with HepQ/B buffer (20mM Tris pH7.5, 1mM TCEP) plus 100mM NaCl. Heparin-binding proteins were eluted with a gradient from 100mM to 1M NaCl over 30 minutes, monitored by absorbance at 280nm. 2ml fractions containing the highest protein concentrations were pooled and applied to a MonoQ anion exchange chromatography column equilibrated with HepQ/B buffer plus 100mM NaCl. Proteins were eluted with a gradient from 100mM to 1M NaCl over 30 minutes, with protein absorbance monitored at

280nm, and 0.3ml fractions were collected. Fractions containing the highest concentrations of eluted protein were pooled and stored in HepQ/B buffer plus 330nM NaCl, corresponding to the salt concentration at which the protein was eluted from the column. Nanodrop A_{280} and the SPN23F12470 protein's predicted extinction coefficient (118390mol/L/cm) were used to determine that SPN23F12470 had been purified to a concentration of 7.0μM.

2.3.5.9 Electrophoretic mobility shift assay

This was done to investigate whether the wild type SPN23F12470 protein was interacting with BOX repeat regions either side of the *ply* gene.

The intergenic region between *ply* and the neighbouring gene SPN23F19460, containing BOX repeat regions, was amplified using the following primers:

FW: GAGAGGAGAATGCTTGCGAC
RV: TAGGAATCTCCTTTTTTCACATTTTAATCTTTC

This yields a product size of 882bp. A region of the *ply* gene was simultaneously amplified using the primers below, also giving a product size of 882bp for use as a control:

FW: ATGGCAAATAAAGCAGTAAATGACTTTATAC
RV: GCCCCCTAAAATAACCGCCTTC

A 10X reaction buffer was prepared containing 200mM Tris (pH 8), 500mM sodium chloride, 10mM dithiothreitol and 30% glycerol. A 1X protein dilution buffer was prepared containing 20mM Tris (pH 8), 200mM sodium chloride, 1mM Tris(2-carboxyethyl)phosphine (TCEP) and 10% glycerol. SPN23F12470 protein was serially diluted in 1X protein dilution buffer to give 5X stock concentrations of 5μM, 2.5μM, 1.25μM, 0.625μM, 0.3125μM and 0.15625μM. Six corresponding reaction mixes were then prepared for each DNA sample containing 1μl 10X reaction buffer, 2μl 5X protein stock, DNA added to a final concentration of 10nM and water to make the reaction volume up to 10μl. A seventh control reaction was also included containing 2μl protein dilution buffer in place of 5X protein stock. The resulting protein concentrations in the final reaction mixes were: 1μM, 500nM, 250nM, 125nM, 62.5nM, 31.25nM and no protein. Samples were then run on a 1.5% agarose gel in 1X TAE buffer for 110 minutes at 90V. Following this the gel was stained in TAE containing 1X SYBR Safe DNA gel stain for 30 minutes, and bands were visualised using a Typhoon FLA 9500.

A second EMSA was run exactly as described above but using a more focused range of protein concentrations from 2.5-80nM. The DNA concentration was also halved to 5nM for this run to aid clearer visualisation of bands.

Radiolabelling of DNA for follow-up EMSAs was done using gamma-32P-ATP by Mark Dillingham. A third 882bp PCR product was amplified for this experiment containing the BOX repeat regions upstream of *ply*, using the following primer sequences:

FW: CAGTAAGATACAGACTGCAAGGC
RV: TGAAAAAGAAACCATTCAGGGGTATTCGG

50nM of each PCR product was added to a reaction containing 2µl phosphorous-32 (gamma-32P-ATP), 2µl T4 polynucleotide kinase (PNK) and 5µl PNK buffer, made up to 50µl with water. This reaction mix was incubated at 37°C for 1 hour, then at 65°C for 20 minutes to heat inactivate the PNK enzyme. Radiolabelled DNA was purified using Microspin S-400 columns as per the manufacturers protocol. 2nM DNA was used in an EMSA exactly as described above, using the protein concentrations from 2.5-80nM, with the exception that the gel was not stained with SYBR safe, but squashed for three hours before being dried using a vacuum-heated gel dryer (BioRad). Bands were visualised as before using a Typhoon FLA 9500.

2.3.6 Statistical comparisons

Statistical comparisons for phenotypic assays were performed using a Student's T-test in Microsoft Excel, unless otherwise stated comparisons were two-tailed and assumed unequal variance.

2.4 Experiments used for both bacterial species

2.4.1 DNA extractions

All DNA extractions were done using High Pure PCR Template Preparation Kit (Roche) as per the manufacturer's instructions. Resulting DNA samples were stored at -20°C

2.4.2 PCR purifications and gel extractions

All PCR purifications were done using the GeneJET PCR purification kit, and gel extractions were done using the GeneJET Gel Extraction kit.

2.4.3 Plasmid extractions

All plasmid extractions were done using the GeneJET Plasmid Miniprep kit.

2.4.4 DNA agarose gel electrophoresis

All PCR products were run on a 0.5-1.5% agarose gel made up using either 1X TAE or 1X TBE. SYBR Safe stain was added to each gel (final concentration 1X) and samples were run for 50-75 minutes at 90V. Either a 1kb or 50bp GeneRuler was used as a size marker. Bands were visualised on a UV transilluminator.

3. Cytolytic toxin production by *Staphylococcus aureus* is dependent upon the activity of the protoheme IX farnesyltransferase

Publication reference

Stevens, E., Laabei, M., Gardner, S., Somerville, G.A. & Massey, R.C. Cytolytic toxin production by *Staphylococcus aureus* is dependent upon the activity of the protoheme IX farnesyltransferase. *Scientific Reports*, 7(13744).

Author contributions

ES carried out the experiments, with the exception of those presented in figures 3.1 and 3.5. Figure 3.1 shows data obtained by ML for a previous publication that led to identification of the *cyoE* gene, and figure 3.5 shows data obtained by SG and GAS due to resources and expertise being more readily available in the Somerville lab at that point. All figures were produced by ES even where data was obtained by collaborators.

3.1 Introduction

The work presented in this chapter constituted my first six-month rotation project as part of my doctoral training programme and was subsequently published in *Scientific Reports*. The successful identification and characterisation of a novel toxicity-associated locus in *S. aureus* led to expansion of the functional genomics approach to a clinical collection of *Streptococcus pneumoniae* isolates, which was the core focus for the remainder of my PhD.

Following a genome-wide association study (GWAS) using a large collection of clinical isolates of the *Staphylococcus aureus* sequence type ST239 (Laabei et al., 2014), several genes of interest were identified as being statistically significantly associated with the toxic phenotype in this bacterium. True positive associations amongst these genes were subsequently identified using corresponding mutants from the Nebraska Transposon Mutant Library in the genetic background USA300 JE2, which were tested for toxicity using the human monocytic cell line THP-1 (see Materials & Methods, section 2.2.2.2 for details). From these assays, disruption of the *cyoE* gene was shown to cause a significant

reduction in toxicity in comparison to the isogenic wild type. To complete this functional verification of the statistical association between *cyoE* and toxicity, work was undertaken to characterise the specific role of this gene in production and secretion of staphylococcal toxins.

This work represents the final stages of the GWAS method and demonstrates that it can be successfully applied to identification of novel contributors to a bacterial phenotype. It builds on the work of Laabei et al. (2014) by taking one locus of interest and focussing on its specific role in relation to toxicity and adds to the work of Yokoyama et al. (2018) which characterised a different locus of interest that arose from this GWAS. By characterising an individual genetic locus to establish its contribution to a phenotype, this work is a small step towards improving our overall understanding of the complex network of genetic factors which contribute to toxicity, and in turn virulence, in *S. aureus*.

Central to *S. aureus*' ability to cause disease is the synthesis of numerous virulence determinants that help facilitate nutrient acquisition and immune evasion. Specifically, *S. aureus* produces adhesins that allow it to adhere to and colonise host tissues; proteins and a capsule that facilitate evasion of the host immune system; and secreted toxins that damage host cells and release nutrients. The expression of these virulence determinants is in part regulated by the Agr quorum sensing system, with a poorly defined post-exponential growth phase signal independent of Agr (Vandenesch et al., 1991; Geiger et al., 2008). The tricarboxylic acid (TCA) cycle activity (Somerville et al., 2002) is also critical for cytolytic activity where TCA cycle mutants have decreased synthesis of secreted toxins. In staphylococci, TCA cycle activity is induced during the post-exponential growth phase, providing the bacterium with carbon, energy, and reduced dinucleotides (Somerville et al., 2003a). TCA cycle-derived four- and five-carbon intermediates are used in biosynthetic reactions to synthesize precursors (e.g. amino acids); while energy (e.g. ATP) drives many critical cellular processes and reduced dinucleotides can donate electrons to the electron transport chain to generate ATP by oxidative phosphorylation.

Recently genetic polymorphisms were analysed within *S. aureus* clinical isolates from the major hospital-associated MRSA lineage ST239 (Holden et al., 2010), and it was found that there was little variability in their adhesive capabilities, but significant variation in cytolytic toxin production (Laabei et al., 2014). The application of GWAS (genome wide association studies) to this data identified a

number of polymorphic loci that were statistically associated with altering cytolytic toxin synthesis. One of those loci, *cyoE*, encodes the enzyme protoheme IX farnesyltransferase, an enzyme involved in catalysing the conversion of haem B to haem O (Saiki et al., 1992; 1993). Haem O is incorporated into the electron transport chain as an electron acceptor, facilitating aerobic respiration and energy production (Hill et al., 1992; Nakamura et al., 1997). As mentioned, TCA cycle mutants have decreased accumulation of cytolytic toxins; hence, the connection between the electron transport chain and the TCA cycle suggests these variants may have decreased TCA cycle activity and toxin accumulation. To address these possibilities, the metabolism and phenotype of a *cyoE*-deficient mutant and complemented strains were analysed.

3.2 Results

3.2.1 Association between toxicity and the polymorphic *cyoE* gene

A recent GWAS study identified an association between a polymorphic version of the *cyoE* gene with changes in the cytolytic activity (toxicity) of *S. aureus*, suggesting this locus may contribute to toxicity (Laabei et al., 2014). A comparison of these data has demonstrated that there was a 2.9-fold reduction (figure 3.1; p=0.0008) in the mean toxicity of *S. aureus* containing the SNP in the *cyoE* gene relative to those with the *cyoE* gene with no SNPs (i.e. that found in the reference strain of the ST239 lineage, TW20).

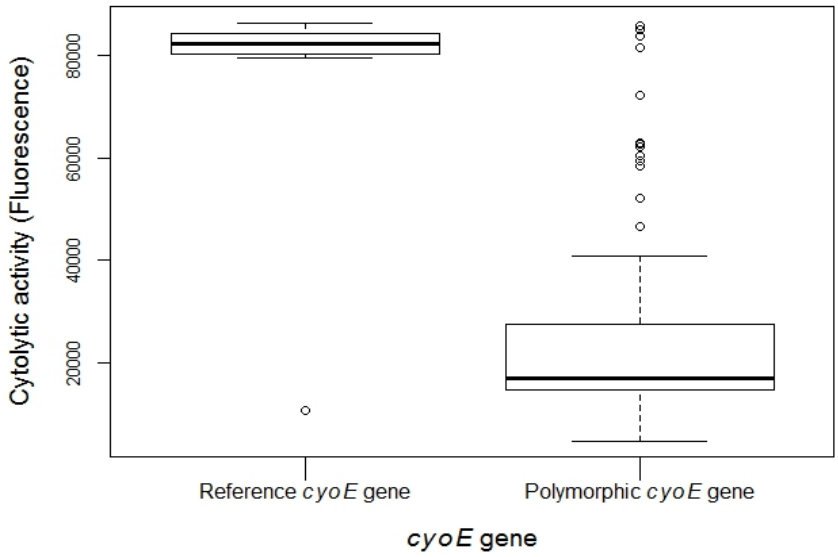


Figure 3.1. Polymorphisms in the *cyoE* gene are associated with a reduction in toxicity. Result of a GWAS study on a collection of 90 ST239 MRSA isolates that associated polymorphisms in the *cyoE* gene with the toxicity of the bacteria. The data presented is the mean toxicity, as measured by lysis of fluorescent dye containing vesicles (Thet et al., 2013) of the isolates with and without polymorphisms in the gene. Eight isolates contained the TW20 reference *cyoE* gene, and 82 isolates contained the SNP. **Data presented here are from vesicle lysis assays originally performed by Maisem Laabei (Laabei et al., 2014).**

The SNP change observed in the collection of clinical strains confers a change from proline to leucine at position 87 in the translated product of the *cyoE* gene. To determine the likelihood of this change affecting the activity of this protein a model of it was built using SWISS-MODEL (Biasini et al., 2014). This model was then viewed in Chimera to visualise where the SNP change occurred and how it might therefore affect toxicity in the clinical strains (figure 3.2a & 3.2b). As this is a membrane protein, a model was also generated using Protter (Omasits et al., 2014) shown in figure 3.2c. Based on these it was hypothesised that the change to leucine could give the loop in which it is located more flexibility and make it more hydrophobic, which could result in the loop flipping inwards on itself. The change also appears to be within a region that could be the haem binding site of this enzyme; figure 3.2b shows the suggested structure of this protein with coulorimbic surface colouring to show positively and negatively charged regions of the protein. The blue region in the lower middle of the structure is thought to be the active site of this enzyme, thus any change to the structure of the loop beneath, particularly the loop flipping inwards, could affect the function of the

active site. Further analysis of the structure of this protein would be required to confirm this hypothesis.

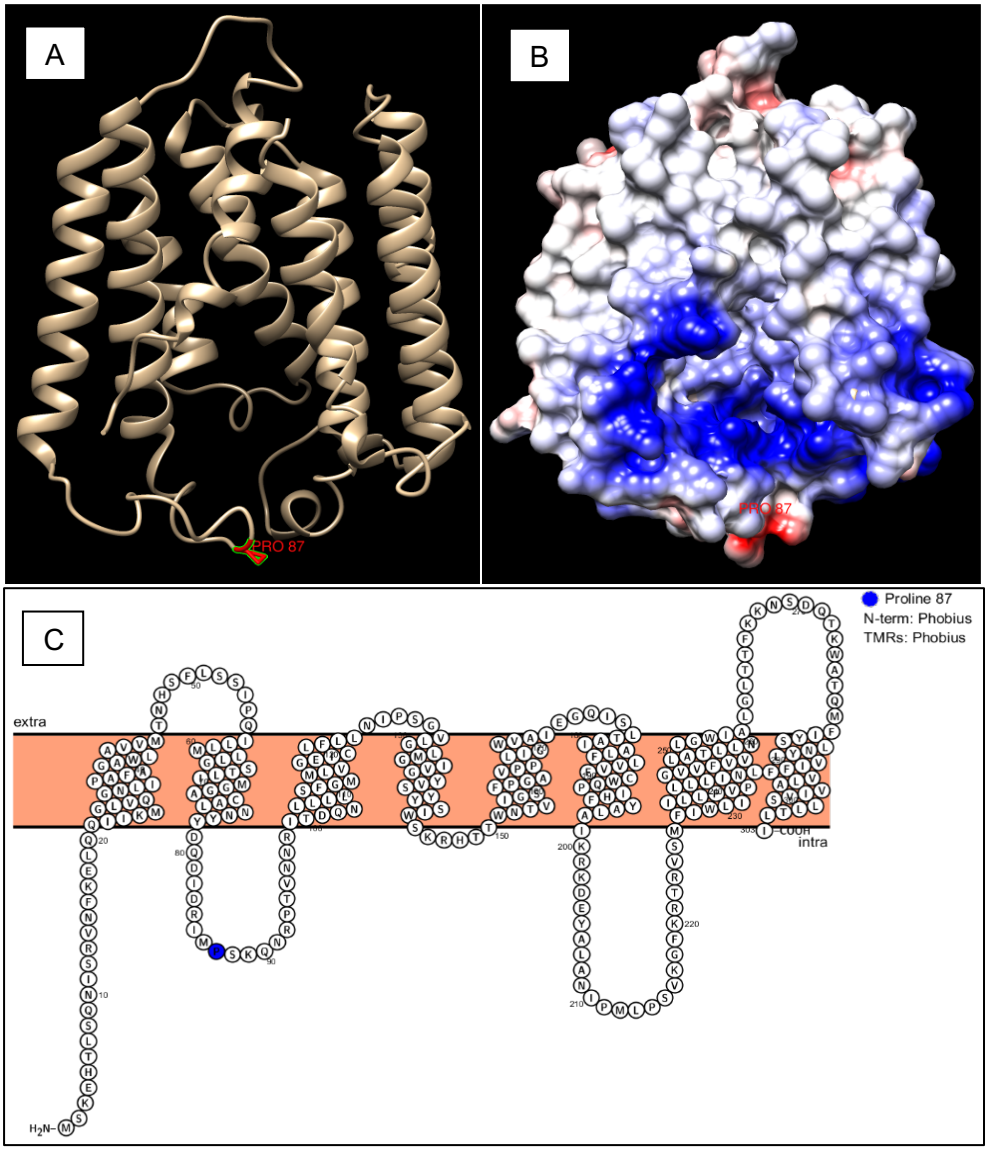


Figure 3.2. Amino acid change from proline to leucine in the polymorphic CyoE protein. A) The position of the amino acid change P87L in the protein, protoheme IX farnesyltransferase, encoded by the *cyoE* gene has been mapped to a ribbon model of the structure of this enzyme, and in **B)** coulorimbic surface colouring has been used to illustrate the white hydrophobic domains of the protein, the blue positively charged amino acid residues, and the red negatively charged amino acid residues. **C)** A model of the protoheme IX farnesyltransferase situated in the cell membrane, and position of the toxicity associated amino acid change, as modelled by Protter (Osmasits et al., 2014).

3.2.2 Functional verification of the contribution protoheme IX farnesyltransferase makes to *S. aureus* toxicity

To verify the association between the *cyoE* gene and toxicity, the cytolytic activity of a Δ*cyoE* transposon mutant from the Nebraska Transposon Mutant library and the isogenic strain JE2 were compared. To quantify cytolytic activity, the THP-1 monocytic cell line, which is sensitive to both the Phenol Soluble Modulins (PSMs) and the other cytolytic toxins secreted by *S. aureus*, was exposed to culture supernatants and toxicity was assessed; concentration of the toxin was not directly quantified. As suggested by the GWAS results, the Δ*cyoE* mutant had significantly decreased toxicity relative to the isogenic wild-type strain (figure 3.3; $p<0.0001$). Complementation of the Δ*cyoE* mutant resulted in restoration of toxicity to wild type levels. To ascertain if the effect of inactivating *cyoE* was strain-dependent, the mutation was transduced into *S. aureus* strain SH1000 and toxicity was assessed. Like in the JE2 background, inactivation of the *cyoE* gene in strain SH1000 caused a significant loss of toxicity (figure 3.3; $p=0.004$). Taken together, these data confirm that the *cyoE* gene contributes to the ability of *S. aureus* to produce toxins.

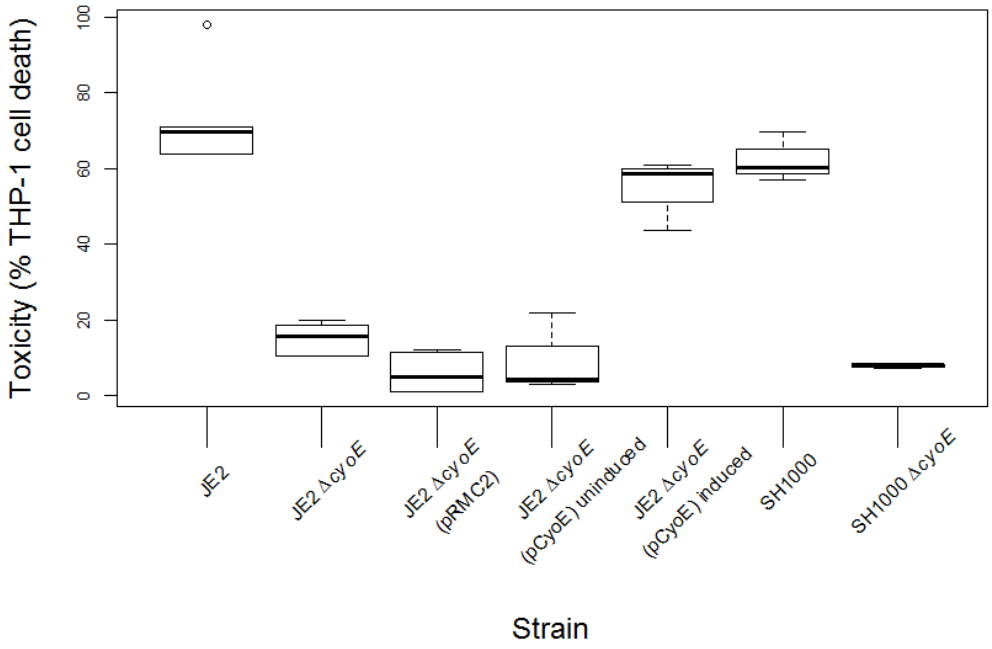


Figure 3.3. Functional verification of the contribution the *cyoE* gene makes to toxicity. The *cyoE* gene was inactivated by transposon insertion in both the JE2 and SH1000 backgrounds. In the JE2 background the loss of toxicity was restored by expressing the gene from an inducible promoter on the pRMC2 vector plasmid. Complementation controls of the empty vector (JE2 Δ*cyoE* (pRMC2)), and the vector with the *cyoE* gene cloned, but uninduced (JE2 Δ*cyoE* (pCyoE)) have been included.

3.2.3 Protoheme IX farnesyltransferase activity affects the ability to activate the Agr quorum sensing system

As the Agr quorum sensing system is a major regulator of toxin synthesis (Recsei et al., 1986; Janzon et al., 1990; Arvidson & Tegmark, 2001) and inactivation of *cyoE* dramatically affects toxicity, it was hypothesised that this effect of the loss of *cyoE* may be mediated through the Agr system. To test this hypothesis, an RNAIII::*gfp* fusion plasmid, which acts as a reporter of Agr activity, was introduced into the JE2 wild-type and Δ*cyoE* mutant strains and fluorescence was monitored over time (figures 3.4a & 3.4b). Optical density data is shown here to demonstrate that inactivation of the *cyoE* gene has a minimal effect on growth of *S. aureus* during this time frame *in vitro*. Growth of the Agr+ strain RN6390B is slower than JE2 and the Δ*cyoE* mutant, which may account for the lower level of RNAIII activity in this strain compared to JE2 as observed in figure 3.4b, however, it is included as a control along with the Agr- strain RN6911 for comparison.

There was a significant reduction in fluorescence in the Δ*cyoE* mutant ($p=0.025$), demonstrating that RNAIII transcription and consequently Agr activation is altered in the Δ*cyoE* mutant relative to the wild-type strain. To further examine the effect of the loss of the *cyoE* gene on Agr activity the expression of toxins known to be under its regulation was quantified, where the secretion of alpha toxin was quantified by western blotting and PSMs by butanol extraction. These assays were performed in triplicate (figure 3.4c) where it was found that on average there was 2.2-fold more alpha toxin, and 1.5-fold more PSMs expressed by the wild type JE2 strains when compared to the Δ*cyoE* mutant. These results confirm that in the absence of the *cyoE* gene the expression and activity of the Agr quorum sensing system is repressed.

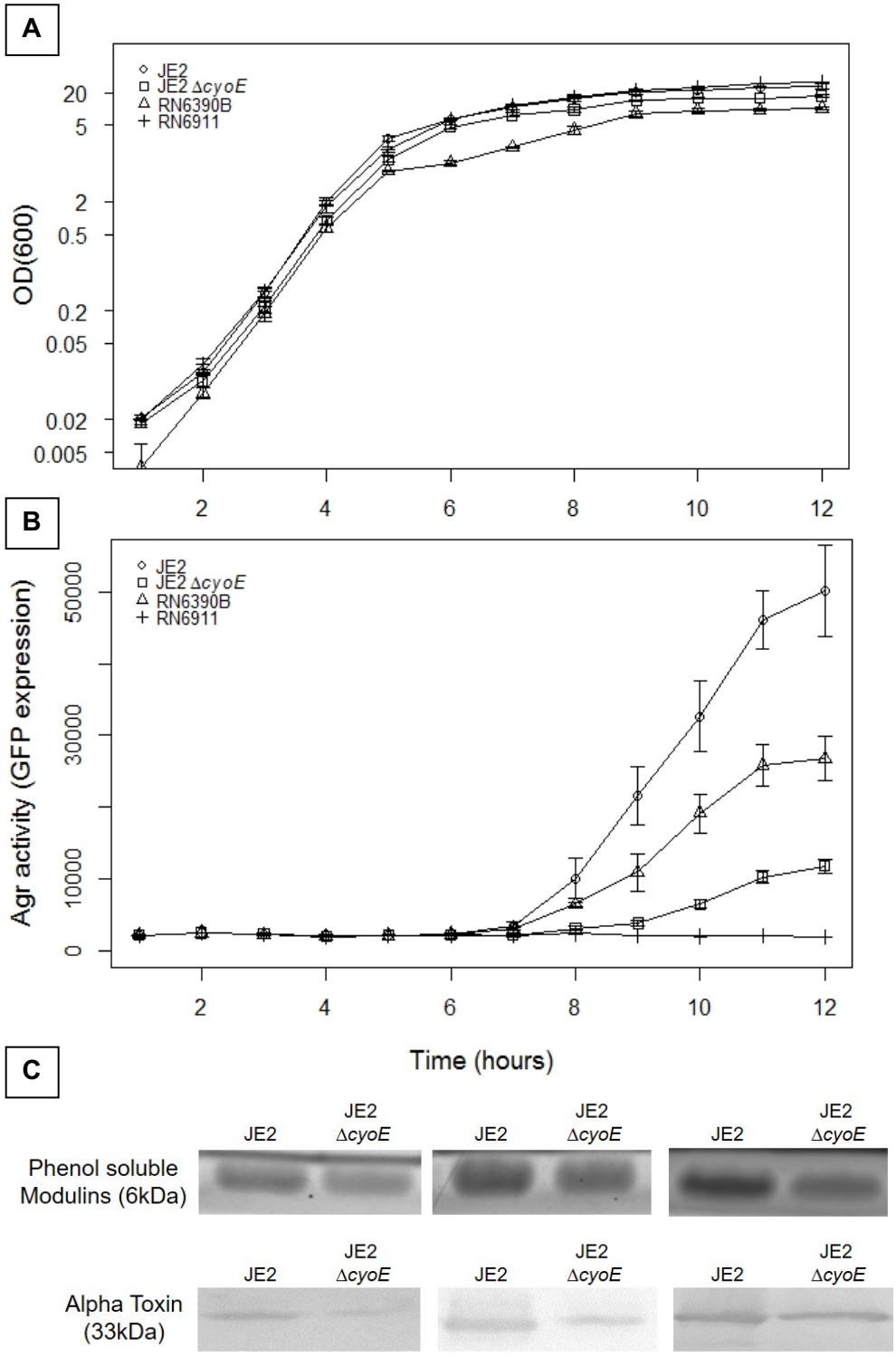


Figure 3.4. Activation of the Agr quorum sensing system is dependent upon protoheme IX farnesyltransferase activity. A) The growth of strains carrying the RNAIII reporter plasmid was monitored over 12 hours demonstrating that inactivation of the *cyoE* gene has minimal effect on the growth of *S. aureus* during this time frame *in vitro*. **B)** Expression of Agr was monitored over 12hours of growth where the inactivation of *cyoE* was found to have a negative effect on Agr activity. Control strains RN6390B (Agr+) and RN6911 (Agr−) have been included for comparison. **C)** The downstream effect of the loss of Agr activity was verified by comparing the production of alpha toxin and the PSMs by both the wild-type and Δ*cyoE* mutant, where the mutant produced 2.2-fold less alpha toxin and 1.5-fold less PSM.

3.2.4 Protoheme IX farnesyltransferase affects the TCA cycle

The protoheme IX farnesyltransferase (CyoE) is essential for electron transport in many organisms, and electron transport facilitates the oxidation of reduced dinucleotides that are generated from TCA cycle activity. As such the inactivation of *cyoE* in *S. aureus* should dramatically decrease TCA cycle activity. The decreased Agr activity (figure 3.4) coupled with decreased TCA cycle activity could provide a potential explanation for the decreased cytotoxicity observed for the Δ*cyoE* mutant, as it has been previously shown secretion of cytotoxins requires TCA cycle activity (Somerville et al., 2002). Furthermore, this would explain why growth of the Δ*cyoE* mutant is not affected, as shown in figure 3.4a; as toxin production is an energetically costly process, sacrifice of this phenotype in response to TCA cycle downregulation would allow the cell to continue to produce sufficient energy for growth using micro-aerobic or anaerobic metabolism pathways, for which TCA cycle activity is not required.

To test this hypothesis, the catabolism of acetate, which requires TCA cycle activity, was assessed. Acetate accumulates in the culture medium during the exponential growth phase due to the incomplete oxidation of carbohydrates when the TCA cycle is repressed. The catabolism of acetate begins when carbohydrates are depleted and the TCA cycle is de-repressed in the post-exponential growth phase. Loss of a functional TCA cycle will result in a deficiency in acetate catabolism, leading to a build-up of acetate in the culture medium and lowering its pH. As expected, the acidification of the media was equivalent, as was the accumulation of acetate between the wild-type, Δ*cyoE* mutant and complemented strains over the first five hours of growth (figures 3.5a & 3.5b). However, once the bacteria ceased growing exponentially (from 5hr onwards) the Δ*cyoE* mutant failed to alkalinise the culture medium and did not catabolise acetate (figures 3.5a & 3.5b).

To specifically verify the effect of *cyoE* on the TCA cycle activity aconitase activity was also quantified, which is a major TCA cycle enzyme catalysing the interconversion of citrate and isocitrate. During the post-exponential growth phase (6hr) it was found that the wild-type and complemented mutant strains both had significantly more aconitase activity when compared to the Δ*cyoE* mutant (figure 3.5c; $p=0.001$). Taken together, the Δ*cyoE* mutant's inability to catabolise acetate and alkalinise the culture medium, demonstrates that inactivation of *cyoE* blocks TCA cycle activity.

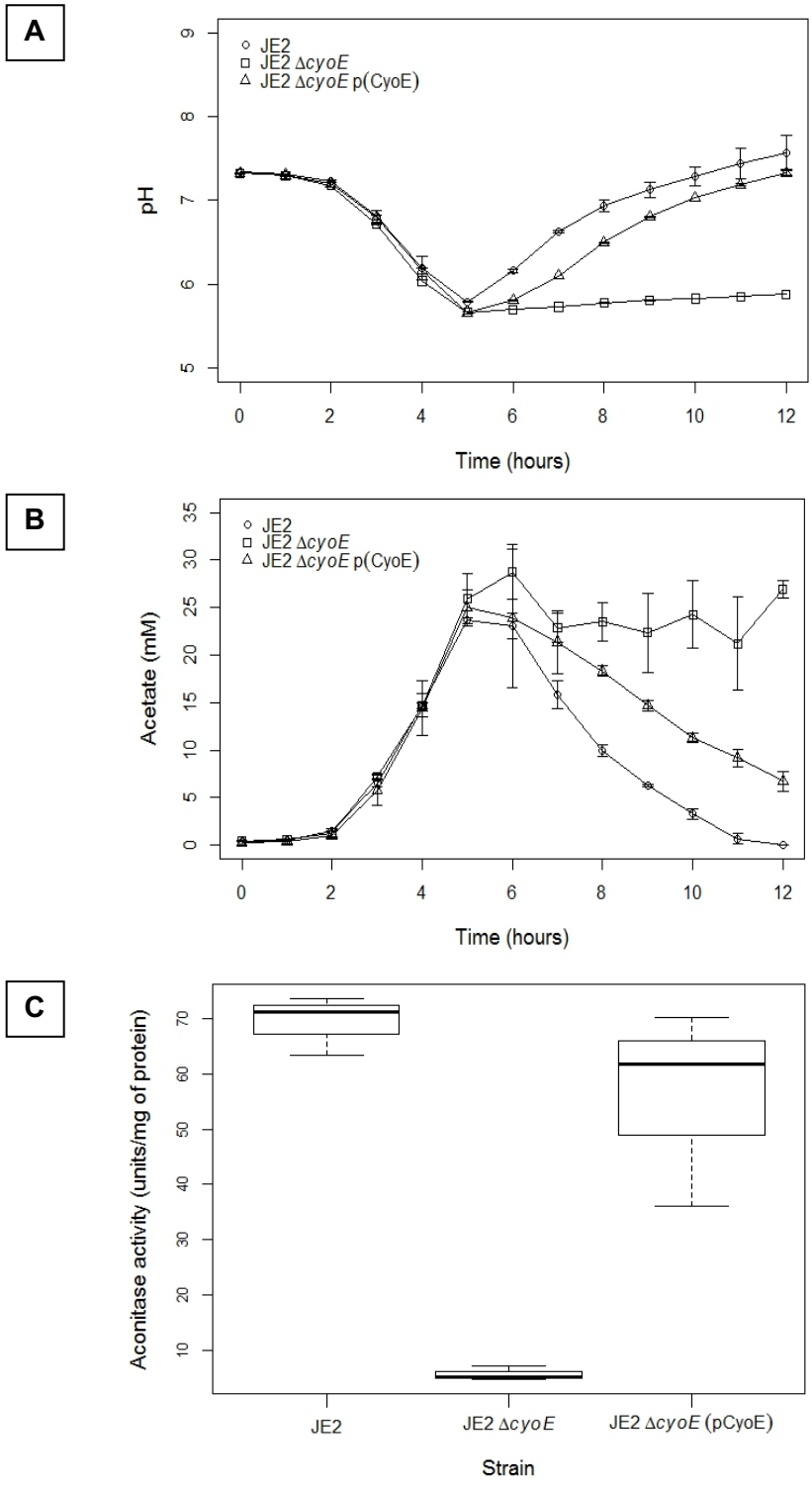


Figure 3.5. Activity of the TCA cycle is affected by the loss of expression of protoheme IX farnesyltransferase. The pH **A)** and acetate **B)** levels of the supernatant of the wild type (JE2), Δ*cyoE* mutant and the complemented mutant were quantified over 12hours of growth. Inactivation of the *cyoE* gene affected both the pH and acetate levels, the effects of which were complemented by expression of the gene from the plasmid pCyoE. **C)** Aconitase activity, a key feature of the TCA cycle, was quantified in the wild type (JE2), Δ*cyoE* mutant and complemented mutant. Inactivation of the *cyoE* gene significantly reduced aconitase activity, the effect of which was complemented by expression of the gene from the pCyoE plasmid. **Data shown here are the results of experiments performed by Stewart Gardner at the University of Nebraska.**

3.2.5 Growth of *S. aureus* under micro-aerobic conditions mimics the effect of a loss of *cyoE*

As an alternative means of demonstrating the effect of reducing TCA cycle activity on toxicity, it was examined whether an effect equivalent to the inactivation of *cyoE* could be achieved by culturing the bacteria in a microaerobic environment. As a facultative anaerobic bacterium, *S. aureus* can grow under anaerobic conditions, however its growth rate is significantly affected with one study finding a nine-fold difference in final cell density when aerobic and anaerobic growth conditions were compared (Coleman, 1985). This study also found that under anaerobic conditions *S. aureus* secretes less alpha toxin, however these effects on cell growth make assessing the relative levels of expression of quorum-sensing dependent proteins complicated. As such, to repress the TCA cycle while minimising the growth defects associated with anaerobic conditions a micro-aerobic environment was created by growing the bacteria in air, but manipulating the flask-to-medium ratio (Somerville & Proctor, 2013). JE2 was grown for 18hrs and the effect of comparing a flask-to-medium ratio of 10:1 and 10:8 on the secretion of toxins was quantified (figure 3.6). The bacteria grown in micro-aerobic conditions had a relatively small (1.9-fold) decrease in biomass when compared to those grown aerobically. There was however a significant effect on toxicity with the bacteria growing in a micro-aerobic environment expressing levels of toxicity equivalent to an aerobically grown Δ*agrB* mutant (figure 3.6).

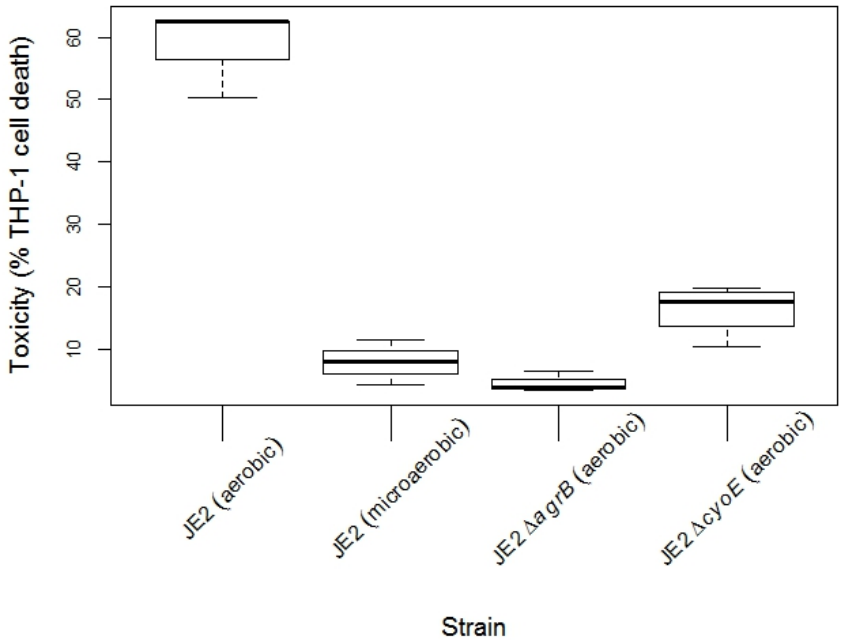


Figure 3.6. Repression of the TCA cycle by growing *S. aureus* under microaerobic conditions repressed the expression of cytolytic toxins. The wild type (JE2) strain was grown under microaerobic (10:8 flask to broth volume ratio) and under aerobic (10:1 flask to broth volume ratio) conditions to determine the effect this has on cytolytic toxin production. An Agr mutant and the Δ*cyoE* mutant both grown aerobically were included. Growth of the wild type strain under microaerobic conditions had an equivalent effect on toxicity as inactivation of either the Agr system or the *cyoE* gene, demonstrating the contribution the TCA cycle makes to the toxicity of *S. aureus*.

3.3 Discussion

S. aureus is one of the most sequenced bacterial pathogens, with many thousands of genomes being publicly available. We are at a point in which this sequence data can be used to understand the biology of this important pathogen in greater detail. This study began as a GWAS analysis of a collection of ST239 isolates and is completed here as the molecular detail of the role of a GWAS identified locus on virulence is characterised. Here it is demonstrated that the *cyoE* gene, which encodes a protoheme IX farnesyltransferase enzyme, plays a critical role in the ability of *S. aureus* to secrete cytolytic toxins. In the Δ*cyoE* mutant, activation of the Agr quorum sensing system is significantly delayed, despite the bacteria reaching sufficient cell density. This effect on the Agr system is coupled to repression of the TCA cycle caused by the loss of haem O production and its role as an electron acceptor in the electron transport chain. This work reiterates the important link between metabolism and virulence in *S. aureus*, but also demonstrates the variability that exists in these attributes amongst clinical isolates causing disease in humans. As genome sequencing

becomes more embedded in clinical diagnostic procedures, information relating to such polymorphic loci could be used to assist in the diagnosis of highly virulent infections.

This study contributes a novel finding to a wider project which aimed to better characterise the genetic basis for toxicity in *Staphylococcus aureus*. Focussing on one locus which was statistically associated through a genome-wide association study with the toxic phenotype in *S. aureus*, it is demonstrated here that the *cyoE* gene plays an important role in cellular metabolism, in turn influencing the production and secretion of cytolytic toxins. The *cyoE* gene encodes the enzyme protoheme IX farnesyltransferase, which processes haem B into haem O for use as an electron acceptor in the electron transport chain. In a mutant deficient in the *cyoE* gene, toxicity is significantly reduced, Agr activity is downregulated and metabolism of the cell, specifically TCA cycle activity, is affected such that the bacterium grows as if it were in micro-aerobic conditions.

Agr activity is critical to the production and secretion of staphylococcal toxins, so it is not surprising that its activity is observed to be downregulated in this toxicity-deficient mutant. Toxin production is an energetically costly process, and the *cyoE* gene is an important component in electron transport for generation of ATP. Without this gene, aerobic respiration cannot continue and the mutant is forced to respire micro-aerobically. Under these conditions, toxin production becomes too energetically costly for the cell, so Agr activity, and in turn toxin production is downregulated accordingly.

Tightly linked to this is activity of the TCA cycle, which is demonstrated to be repressed in the *cyoE* mutant. This repression is caused by the loss of function of the electron transport chain. This deficiency in electron transport means reduced dinucleotides (i.e. NADH/$FADH_2$) produced by the TCA cycle cannot be oxidised, and the TCA cycle becomes inhibited through negative feedback. As a result, metabolic by-products of exponential phase growth, which would normally be catabolised by the TCA cycle during stationary phase, build up in the culture medium and the mutant behaves as if it were growing in micro-aerobic conditions.

Altogether the effect of the *cyoE* mutation is to cause a deficiency in electron transport which prevents the mutant from respiring aerobically. This has a critical effect on other metabolic pathways and reduces the amount of ATP that the cell

can produce. In response to these metabolic changes, activity of the regulatory Agr system is repressed and toxin production is significantly reduced.

While this work focusses on a very specific contributor to toxin production in *S. aureus*, it builds on a wider body of knowledge regarding the complicated genetic regulatory network that influences staphylococcal toxicity. In this era of antibiotic resistance, and with the increasing availability of genetic information, it will become increasingly important to understand the intricacies of how bacteria such as *S. aureus* are able to cause disease if we are to find alternative treatment and prevention methods, and this can be achieved through better understanding individual gene function in relation to disease.

It is shown here that GWAS can be a useful tool in working towards this goal. Using a statistical approach along with whole genome sequence data, individual genetic contributors to a phenotype can be identified and characterised to establish their specific function in the context of bacterial virulence. This work represents the later stages of this process, and demonstrates how understanding individual gene function can help to further our understanding of how bacteria cause disease.

4. Investigating the genetic basis for toxicity in *Streptococcus pneumoniae* using genome-wide association approaches

4.1 Introduction

Streptococcus pneumoniae is a major cause of disease worldwide, ranging from minor infections such as sinusitis and otitis media, to serious invasive disease such as pneumonia, meningitis and septicaemia. A key virulence factor of the pneumococcus is the cytolytic pore-forming toxin pneumolysin, which can cause extensive damage to host tissues at the site of infection. Following successful identification of novel toxicity-associated loci in *Staphylococcus aureus* using genome-wide association (Laabei et al., 2014; 2015), and the means by which these loci affect toxin production (Stevens et al., 2017; Yokoyama et al., 2018), this functional genomics approach was applied to the investigation of toxicity-affecting loci in *S. pneumoniae*.

The aims of this project were:

1) To identify novel genetic contributors to toxicity in *S. pneumoniae* to further our understanding of how this bacterium causes disease
2) To compare results from three different GWAS methods to aid prioritisation of genes for functional verification

A crucial difference between *S. aureus* and *S. pneumoniae* is that *S. pneumoniae* produces only one toxin, while *S. aureus* produces up to fifteen. However, the genetic regulation of toxicity is complex and incompletely understood. By further elucidating genetic effectors of toxicity in this organism, this work contributes to the increasing body of knowledge regarding how *S. pneumoniae* causes disease, ultimately working towards better treatment and prevention of pneumococcal disease. From the functional genomics approach used here, a novel effector of toxicity is identified in *S. pneumoniae*.

A further aspect of this project is to extend the functional genomics approach used in previous work on *S. aureus* by applying three different GWAS methods to the dataset. The purpose of this is to aid prioritisation of statistically significant loci for further investigation, to determine whether true positive results are identified consistently between the different methods.

4.2 Results

4.2.1 Generation of the dataset for GWAS

GWAS requires a quantitative set of phenotypic data and a corresponding set of genotypic data from a large number of bacterial isolates. In this case, the phenotypic dataset consists of toxicity data from a collection of 166 clinical isolates of the clonal complex PMEN-1, taken from patients with pneumococcal disease, each of which varies based on single nucleotide polymorphisms present throughout the genome. These single nucleotide polymorphisms make up the genotypic dataset. Clinical isolates and their corresponding SNP data were obtained from Stephen Bentley and Nicholas Croucher at the Sanger Institute, Cambridge, with technical assistance from Rebecca Gladstone.

4.2.1.1 Toxicity assays for clinical strains

A haemolytic activity assay was developed using *S. pneumoniae* TIGR4 wild type and an isogenic TIGR4 pneumolysin-deficient strain, obtained from Angela Nobbs at the University of Bristol, to quantify variations in toxicity across the collection of clinical isolates. Use of a pneumolysin-deficient mutant as a control demonstrated that lack of the toxin gene *ply* resulted in no lysis of red blood cells, therefore showing that pneumolysin alone was responsible for cell lysis, and any variations in toxicity of the clinical isolates could be attributed to variations in pneumolysin production (see figure 4.1).

For this assay, whole cell culture at mid-exponential phase was serially diluted 4-fold, 6 times in a 96-well plate and mixed with 2% sheep red blood cells. Cells were incubated in this manner for 1 hour at 37°C with 5% CO_2, and lysis of red blood cells was measured by transferring the supernatant to a fresh plate and reading absorbance at 596nm (see Materials & Methods, section 2.3.2 for further details).

The wild type TIGR4 and pneumolysin-deficient mutant were included as controls in every assay. Average haemolytic activity for each control along with a buffer-

only blank is shown in figure 4.1 below, with toxicity shown as absorbance of the supernatant at 596nm; high absorbance values indicate high toxicity, and vice versa. All toxicity data presented is for the first 4-fold dilution, which was determined to be the most useful for showing variations in toxicity.

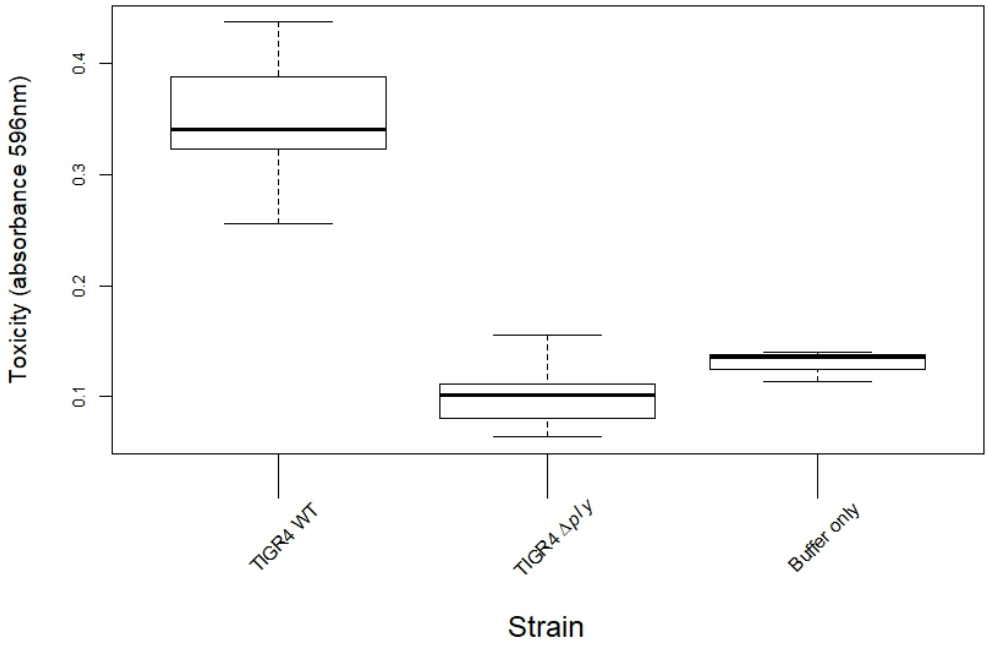


Figure 4.1: **Haemolytic activity of wild type TIGR4 and isogenic TIGR4 Δ*ply* strains.** These were used as controls on each 96 well plate in the subsequent haemolytic activity assays. For TIGR4 WT and TIGR4 Δ*ply* strains n=17, for the buffer only n=3. Boxes represent the interquartile range of the dataset, with the thick black line representing the median. Whiskers show the highest and lowest data points. The difference between the Δ*ply* strain and the buffer is statistically significant in a T-test (p=0.04), however there are only 3 data points for the buffer, compared to 17 for Δ*ply*, and there is also a greater amount of variation in the Δ*ply* dataset as this represents data from several assays combined.

Each of the 166 clinical isolates was then run through the haemolytic activity assay to gather the quantitative phenotypic dataset, the results of which are shown in figure 4.2. Data are presented as a scatterplot in descending order of toxicity for simplicity.

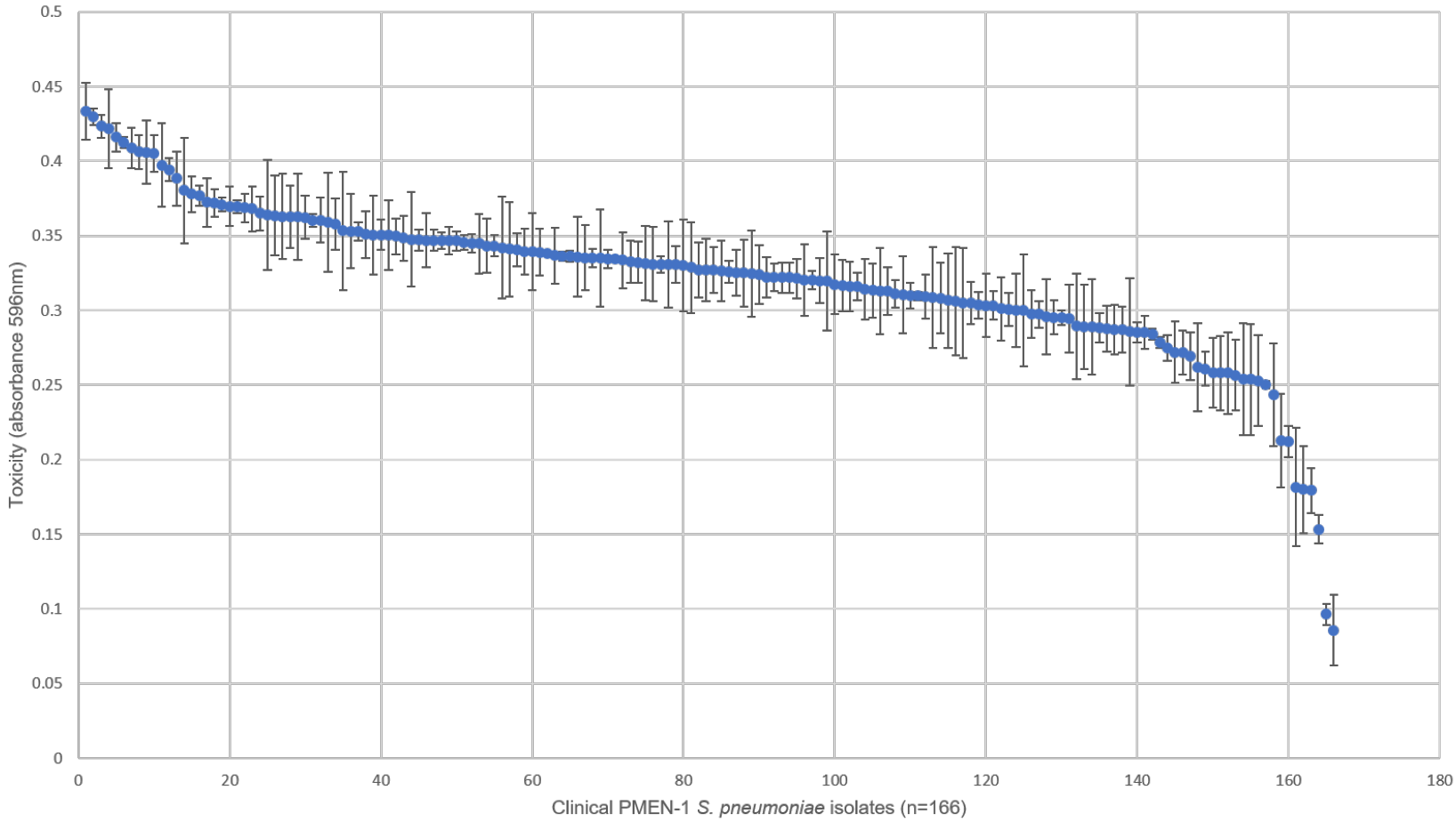


Figure 4.2. Haemolytic activity of 166 isolates from the PMEN-1 clinical collection of strains. Values given are the averages of three biological repeats for each strain. Error bars represent standard error of the mean.

4.2.2 Genome-wide association studies

The phenotypic data generated from the toxicity assays and genome sequence data for each clinical isolate were then used in three different GWAS methods:

1) SNP-based GWAS – run with assistance from Mario Recker at the University of Exeter
2) SEER – run by John Lees, formerly at the Sanger Institute, Cambridge
3) Bugwas – run by Daniel Wilson and Sarah Earle at the University of Oxford

4.2.2.1 SNP-based GWAS

With guidance and R scripts from Dr. Mario Recker, my co-supervisor based at the University of Exeter, the SNP data file was converted into numerical values, with 1 representing presence of a SNP at a given position, and 0 representing no SNP change at a given position, relative to the reference genome of ATCC700669 (SPN23F). Missing data was converted to NA values and therefore excluded from statistical analysis. Positions with a high number of missing values were also excluded; this is the genotype frequency, the threshold for which was set at 5%; this means that any SNP positions which had 95% or more NA values were excluded from the study. A further threshold was set for positions with less than 3% SNP changes; this is the minor allele frequency and excludes SNP positions at which fewer than 3% of isolates had a SNP change. Pearson's product-moment correlation was then performed using the filtered SNP data and toxicity data from each strain, to identify associations between SNP positions and the toxic phenotype.

Following this, a permutation test was conducted to determine whether the observed results are exchangeable under the null hypothesis: this being that there is no significant association between a SNP position and the toxic phenotype. A permutation test illustrates that significant results are not obtained through chance but are actually correlated with the phenotypic data. A permutation test is done by scrambling the dataset to remove the association between genotype and phenotype, and then re-doing the correlation analysis to see if the same significant results are observed. This is done, for example, a thousand times, therefore scrambling results which are biologically significant and showing the likelihood of specific SNP positions remaining significant once the dataset has been scrambled numerous times. The results of this permutation are

used to produce empirically derived p-values which gives more support to the statistically significant results produced from the GWAS. Following permutation, p-values were converted to -log10 and plotted on a Manhattan plot, as shown in figure 4.3 below. The threshold for statistical significance was set to p=0.05, results which cross this threshold are highlighted in red in figure 4.3.

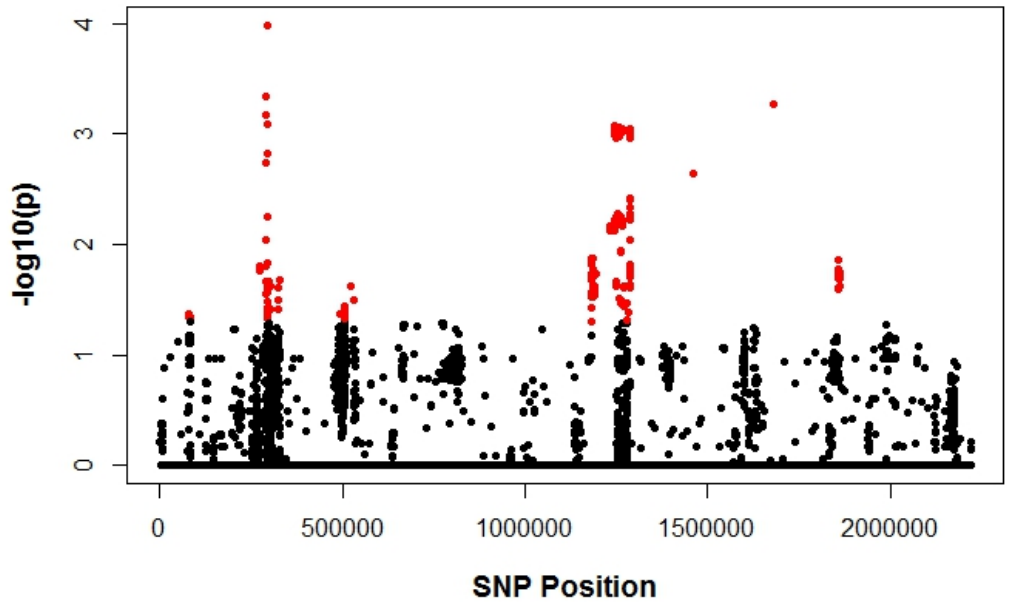


Figure 4.3: Manhattan plot showing GWAS results. Thresholds were set to eliminate data with a genotype frequency of less than 5% and a minor allele frequency of less than 3%. P-values shown were empirically derived following permutation of the dataset, and statistically significant values have been colour coded in red based on a threshold for significance which is set as –log10(0.05).

Loci identified using this method as statistically significantly associated with toxicity totalled 548 within 56 genes and 11 intergenic regions. These genes are listed in table 4.1 below.

Table 4.1: **SNP-based GWAS results**. Genes in reference strain ATCC 700669 (SPN23F) in which SNPs were identified that were statistically significantly associated with toxicity using SNP-based GWAS.

Gene Name	Gene Function
agaS	putative tagatose-6-phosphate aldose/ketose isomerase
SPN23F00810	putative aldose 1-epimerase
SPN23F00820	glyoxalase/bleomycin resistance protein/dioxygenase superfamily protein
SPN23F02870	putative glycosylhydrolase
pbpX	putative penicillin binding protein 2x
mraY	phospho-N-acetylmuramoyl-pentapeptide-transferase

Intergenic	intergenic between *mraY* and *clpL* - contains BOX repeats
clpL	putative ATP-dependent protease ATP-binding subunit ClpL
SPN23F03140	conserved hypothetical
Intergenic	intergenic between SPN23F03370 and *aliA* - contains RUP repeat
aliA	putative extracellular oligopeptide-binding protein
SPN23F03400	cell wall surface anchored protein
valS	valyl-tRNA synthetase
Intergenic	intergenic between SPN23F05220 and *pheS* - contains RUP repeat
pheS	phenylalanyl-tRNA synthetase alpha chain
Intergenic	intergenic between SPN23F05240 and *pheT*
pheT	phenylalanyl-tRNA synthetase beta chain
SPN23F05380	putative uncharacterised protein
SPN23F05470	putative amino acid ABC transporter, permease protein
ntpE	V-type sodium ATP synthase subunit E
ntpK	V-type sodium ATP synthase subunit K
SPN23F12160	putative ROK family protein
Intergenic	intergenic between SPN23F12160 and SPN23F12170
SPN23F12170	putative oxidoreductase
Intergenic	intergenic between SPN23F12170 and SPN23F12180
SPN23F12180	putative sialidase (neuraminidase)
SPN23F12190	conserved hypothetical protein
SPN23F12200	putative sodium:solute symporter
nanA	probable N-acetylneuraminate lyase
Intergenic	intergenic between SPN23F12270 and SPN23F12280
SPN23F12640	putative uncharacterised protein
SPN23F12650	putative uncharacterised protein
SPN23F12660	putative uncharacterised protein
SPN23F12670	putative uncharacterised protein
Intergenic	intergenic between SPN23F12670 and SPN23F12680
SPN23F12680	putative uncharacterised protein
SPN23F12760	putative membrane protein
SPN23F12770	putative uncharacterized protein
Intergenic	intergenic between SPN23F12770 and SPN23F12780
SPN23F12780	putative conjugative transposon DNA recombination protein
SPN23F12820	putative uncharacterized protein (pseudogene)
SPN23F12840	putative conjugal transfer protein (possible involvement with cell wall metabolism)
SPN23F12850	putative conjugal transfer protein
SPN23F12860	putative uncharacterized protein
SPN23F12870	putative uncharacterized protein
SPN23F12880	putative uncharacterized protein

traG	putative conjugal transfer protein TraG
SPN23F12900	putative uncharacterized protein
SPN23F12910	putative uncharacterized protein
SPN23F12930	conserved hypothetical protein
orf21	conjugative transposon FtsK/SpoIIIE-family protein
tetM	conjugative transposon tetracycline resistance protein
SPN23F13110	putative CAAX amino terminal protease
SPN23F13120	putative uncharacterized protein
SPN23F13130	putative uncharacterized protein
SPN23F13140	putative uncharacterized protein
SPN23F13150	putative DNA methylase
Intergenic	intergenic between SPN23F13150 and SPN23F13160
SPN23F13160	putative replication initiator protein
SPN23F15080	putative membrane protein
priA	putative primosomal protein N'
SPN23F19110	putative polysaccharide repeat unit polymerase
SPN23F19120	sucrose phosphorylase
Intergenic	intergenic betwteen SPN23F12190 and *msmG* - contains BOX repeats
msmG	multiple sugar-binding transport system permease protein
msmF	multiple sugar-binding transport system permease protein
msmE	multiple sugar-binding protein precursor

4.2.2.2 SEER

Sequence element enrichment analysis (SEER) is a GWAS method developed by John Lees (Lees et al., 2016) which uses short sections of DNA instead of SNPs as the unit of genetic variation. These ‘sequence elements’, referred to as k-mers, vary in length and are produced from the original genome sequence data instead of SNP data. The benefit of using k-mers in place of SNPs alone is that they capture other types of genetic variation along with individual polymorphisms, including indels, recombinations and variations in promotor sequences and gene content. Further differences between this and the SNP-based GWAS method are that SEER uses logistic and linear regression to identify associations between genotype and phenotype, it controls for population structure using metric multi-dimensional scaling, and corrects for multiple testing using a conservative Bonferroni cut off for significance. It is open to debate as to whether this is better than the SNP-based approach, hence why the different methods are combined in

this project, to see if one method produces more true positive associations than another.

SEER was conducted on this dataset by John Lees and I did not contribute to running this analysis, only to analysing the results. This approach identified 34 genes and 9 intergenic regions containing k-mers which were statistically significantly associated with the toxic phenotype, details of which are presented in table 4.2 below. Four of these genes were also identified as containing statistically significant loci by the SNP-based GWAS method and these are highlighted in red.

Table 4.2: **SEER results.** Genes in reference strain ATCC 700669 (SPN23F) in which k-mers were identified as being statistically significantly associated with toxicity using SEER.

Gene name	Gene function
trcF	putative transcription repair coupling factor
SPN23F00770	sugar phosphotransferase system - sorbose-specific family IIC
SPN23F01610	putative ABC transporter system permease protein
Intergenic	intergenic between SPN23F01610 and SPN23F01620
SPN23F01620	putative membrane protein
SPN23F01630	putative membrane protein
SPN23F01640	putative histidine sensor kinase
SPN23F05060	putative uncharacterised protein
Intergenic	intergenic between SPN23F05060 and SPN23F05070 - contains BOX repeats
SPN23F05070	conserved hypothetical protein
SPN23F06470	putative coenzyme
SPN23F07400	putative IS640-Spn1 transposase (pseudogene)
Intergenic	intergenic between SPN23F07430 and SPN23F07440
SPN23F07440	putative amino acid ABC transporter system permease protein
SPN23F08070	putative type I RM modification system
SPN23F09200	putative IS630-Spn1 transposase (pseudogene)
gidB	methyltransferase GidB
SPN23F12470	putative DNA helicase II UvrD
SPN23F12510	putative NTPase protein
SPN23F12530	putative phosphoserine phosphatase
SPN23F12690	putative lantibiotic transport/processing ATP-binding protein
SPN23F12701	putative lantibiotic precursor
Intergenic	intergenic between SPN23F12701 and SPN23F12710
SPN23F12710	putative lantibiotic ABC transporter
Intergenic	intergenic between SPN23F12820 and SPN23F12830
SPN23F12830	putative uncharacterised protein
SPN23F12840	putative conjugal transfer protein

traG	putative conjugal transfer protein TraG
SPN23F12930	conserved hypothetical protein
SPN23F13710	putative membrane protein
SPN23F15290	autolysin
Intergenic	intergenic between SPN23F15290 and *hol*
hol	antiholin
SPN23F15590	hypothetical phage protein
scrA	putative sucrose-specific phosphotransferase system (PTS), IIABC component
fmt	methionyl-tRNA formyltransferase
Intergenic	intergenic between *fmt* and *priA*
priA	putative primosomal protein N'
SPN23F18300	conserved hypothetical protein
Intergenic	intergenic between trpE and SPN23F18390 - contains BOX repeats
SPN23F19510	putative IS1381 transposase (pseudogene)
Intergenic	intergenic between SPN23F21680 and *pcpA*
pcpA	cell surface choline binding protein PcpA

4.2.2.3 Bugwas

Bugwas is an alternative GWAS method developed by Daniel Wilson and Sarah Earle at the University of Oxford (Earle et al., 2016). This approach uses linear mixed models (LMMs) to identify associations between genotype and phenotype and, like SEER, controls for population structure. However, Bugwas uses principal components to test for lineage effects and control for population structure as opposed to metric multi-dimensional scaling. This approach employs the Bonferroni correction to correct for multiple testing. Here, both SNP data and k-mer data are used as the unit of genetic variation to test for association with the phenotype; results are presented from four different analyses; one using bi-allelic SNPs as the genotypic data (the SNP change is always the same compared to the reference genome), one using tri/tetra-allelic SNPs (the SNP change could be to one of two or three different bases respectively, compared to the reference genome), one using fixed-length k-mers and one using variable-length k-mers.

Bugwas was conducted on this dataset by Sarah Earle and I did not contribute to running this analysis, only to analysing the results. Results from the SNP-based analysis identified 57 genes and 9 intergenic regions containing SNPs which were statistically significantly associated with the toxic phenotype; three of these genes were also identified as statistically significant by

the SNP-based GWAS method (text coloured red), and one gene was identified as statistically significant by SEER (text highlighted in yellow), as shown in table 4.3 below.

Table 4.3. Bugwas SNP results. Genes in reference strain ATCC 700669 (SPN23F) in which SNPs were identified that were statistically significantly associated with toxicity using Bugwas.

Bi-allelic loci	
Gene name	**Gene function**
trcF	putative transcription repair coupling factor
SPN23F00830	conserved hypothetical protein
SPN23F00840	putative uncharacterized protein
SPN23F00850	conserved hypothetical protein
SPN23F01020	binding-protein-dependent transport system membrane protein
agaD	putative N-acetylgalactosamine-specific phosphotransferase system (PTS), IID component
SPN23F03010	heparinase II/III-like protein
SPN23F03020	LacI family regulatory protein
SPN23F03030	membrane protein
Intergenic	intergenic between SPN23F03030 and SPN23F03040
pbpX	penicillin-binding protein 2X
mraY	phospho-N-acetylmuramoyl-pentapeptide- transferase
glnQ3	putative glutamine transporter, ATP-binding protein 3
SPN23F07870	conserved hypothetical protein
Intergenic	intergenic between *lacT2* and *lacD2*
udk	uridine kinase
SPN23F11330	conserved hypothetical protein
ffh	signal recognition particle protein
SPN23F14040	putative transposase remnant (pseudogene)
SPN23F17530	conserved hypothetical protein
SPN23F17550	conserved hypothetical protein
Intergenic	intergenic between SPN23F17550 and SPN23F17560
SPN23F17560	GTP-binding protein
SPN23F17580	CorA-like Mg2+ transporter protein
SPN23F17800	putative glycosyltransferase
psrP	cell wall surface anchored protein
Intergenic	intergenic between SPN23F18050 and SPN23F18060
SPN23F18400	hypothetical protein
SPN23F18970	sugar phosphotransferase system (PTS), IIABC component
SPN23F20870	putative exported protein
dnaN	DNA polymerase III, beta chain

SPN23F00470	putative uncharacterized protein (pseudogene)
purK	putative phosphoribosylaminoimidazole carboxylase ATPase subunit
ribE	riboflavin synthase alpha chain
SPN23F02770	putative CAAX amino terminal protease family membrane protein
folP	dihydropteroate synthase
folC	putative putative folylpolyglutamate synthase
SPN23F03400	cell wall surface anchored protein
alsS	acetolactate synthase large subunit
SPN23F05080	putative uncharacterized protein
SPN23F05210	beta-glucoside-specific phosphotransferase system (PTS), IIABC component
SPN23F05680	putative carboxypeptidase
rplA	50S ribosomal protein L1
Intergenic	intergenic between SPN23F06400 and *thiM* - contains BOX repeats
rpiA	ribose 5-phosphate isomerase A
pfk	6-phosphofructokinase
SPN23F08620	putative replication initiation protein (pseudogene)
Intergenic	intergenic between *comEC* and SPN23F08820
Intergenic	intergenic between *prsA* and SPN23F09070
Intergenic	intergenic between SPN23F09270 and *lmb*
SPN23F10470	putative glycerate kinase-family protein
zmpA	IgA-protease
potD	spermidine/putrescine extracellular binding protein
aldB	alpha-acetolactate decarboxylase
pstS	putative phosphate ABC transporter, extracellular phosphate-binding lipoprotein
pclA	putative NADPH-dependent FMN reductase
nanA	sialidase A (neuraminidase A)
Intergenic	intergenic between SPN23F18610 and *xpt* - contains BOX repeats and SPRITE repeat
SPN23F20830	putative pyrrolidone-carboxylate peptidase
hasC	UTP-glucose-1-phosphate uridylyltransferase
SPN23F21510	putative Major Facilitator Superfamily protein
glpK	glycerol kinase
SPN23F17520	putative isochorismatase
SPN23F17570	putative hydrolase
Tri- or tetra-allelic loci	
SPN23F14660	putative amino acid ABC transporter ATP-binding protein
mraY	phospho-N-acetylmuramoyl-pentapeptide- transferase
pbpX	penicillin-binding protein 2X
SPN23F00840	putative uncharacterized protein

Results from the two k-mer-based analyses identified 6 genes and 1 intergenic region containing k-mers that were statistically significantly associated with toxicity. None of the genes identified using the fixed-length k-mer analysis were identified in any of the other GWAS methods, while 2 genes identified using the variable-length k-mer analysis were also identified by both the SNP-based GWAS method and the SNP-based Bugwas method. These genes are shown in red in table 4.4 below.

Table 4.4. Bugwas k-mer results. Genes in reference strain ATCC 700669 (SPN23F) in which k-mers were identified that were statistically significantly associated with toxicity using bugwas.

Fixed-length k-mers	
Gene name	**Gene function**
SPN23F21490	putative uncharacterized protein
SPN23F20850	MarR family regulatory protein
Intergenic	intergenic region containing BOX-repeat elements
SPN23F20870	putative exported protein
Variable-length k-mers	
pbpX	penicillin-binding protein 2X
mraY	phospho-N-acetylmuramoyl-pentapeptide- transferase
SPN23F05430	two-component system, response regulator

4.2.2.4 Summary of key features and differences between the three methods

SNP-based GWAS:

- Uses SNPs as the unit of genetic variation
- Does not control for population structure
- Corrects for multiple testing using permutation

SEER:

- Uses k-mers as the unit of genetic variation
- Controls for population structure using metric multi-dimensional scaling
- Corrects for multiple testing using the Bonferroni correction

Bugwas

- Uses both SNPs and k-mers as the unit of genetic variation
- Controls for population structure using principal component analysis
- Corrects for multiple testing using the Bonferroni correction

4.2.3 Functional verification of GWAS results

GWAS approaches are notorious for producing false positive results. Direct associations are causative genetic loci that are directly responsible for changes in the phenotype. Indirect associations, or false positives, are genetic loci that are

falsely associated with the phenotype because polymorphisms at these positions have been inherited by chance alongside the causative locus, with which they are in linkage disequilibrium. For this reason, polymorphisms that have been statistically associated with the toxic phenotype tend to cluster around a particular genetic locus as seen in figure 4.3, where only one polymorphism may in fact be causative of the observed phenotype.

To identify which genetic loci are directly associated with this phenotype, it is necessary to verify the function of each gene of interest in the lab. Here, this was done by attempting to disrupt each gene significantly associated with toxicity and testing the toxic phenotype of the resulting mutant using the haemolytic activity assay. Through discussion with collaborator Tim van Opijnen at Boston College, we obtained transposon mutants in 15 associated genes in the *S. pneumoniae* TIGR4 2394 background, which we could use to begin the process of functional verification (results section 4.2.3.1). Alongside this, it was intended that I would make whole-gene deletions of each associated gene in the reference *S. pneumoniae* strain ATCC 700669 (SPN23F), results for which are shown in section 4.2.3.2.

4.2.3.1 Transposon mutants in the TIGR4 2394 background strain

35 transposon mutants were obtained in the TIGR4 2394 background from collaborators Tim van Opijnen and Derek Thibault at Boston College. These mutants have transposon insertions in 15 associated genes, with different mutants having transposon insertions at different positions within the same gene. The Opijnen lab has previously constructed a large number of transposon mutants in the TIGR4 2394 background, and the ones provided here were all of the mutants in the Opijnen lab collection that crossed over with the GWAS results from the SNP-based method. Table 4.5 below gives details of each of these mutants, including how many were provided per gene, and the location of transposon insertions.

Table 4.5: **Transposon mutant details.** Genes in which knockout mutants were provided for us by Professor Tim van Opijnen in the background strain TIGR4 2394.

Gene name in ATCC 700669 (SPN23F)	Gene function	Locus tag in TIGR4	No of mutants	Position(s) of Tn insertion(s)
agaS	putative tagatose-6-phosphate aldose/ketose isomerase	SP0065 - agaS	2	69879, 70282
SPN23F00820	glyoxalase/bleomycin resistance protein/dioxygenase superfamily protein	SP0073	3	79692, 80388, 80391
SPN23F02870	putative glycosylhydrolase	SP0312	1	286193
clpL	putative ATP-dependent protease ATP-binding subunit ClpL	SP0338	1	312171
SPN23F03140	conserved hypothetical	SP0341	2	314176, 314835
aliA	putative extracellular oligopeptide-binding protein	SP0366 - aliA	1	339543
SPN23F03400	cell wall surface anchored protein	SP0368	5	342610, 342722, 343006, 344331, 344446
SPN23F05380	putative uncharacterised protein	SP0595	1	564606
SPN23F12160	putative ROK family protein	SP1324	4	1247417, 1247683, 1247707, 1247993
SPN23F12180	putative sialidase (neuraminidase)	SP1326	5	1249836, 1250124, 1250884, 1250998, 1251193,
SPN23F12190	conserved hypothetical protein	SP1327	1	1251839
SPN23F12200	putative sodium:solute symporter	SP1328	2	1252858, 1253242
nanA	probable N-acetylneuraminate lyase	SP1329	4	1253826, 1253962, 1254265, 1254368
SPN23F15080	putative membrane protein	SP1544 - aspC	1	1451657
SPN23F19120	sucrose phosphorylase	SP1894 - gtfA	2	1800323, 1800573

These mutants were each tested for haemolytic activity to begin functional verification of genes identified from GWAS to verify any true positive associations. The figure below shows toxicity data for each of the TIGR4 2394 mutants and the 2394 wild type strain, based on the averages of ten biological repeats per mutant. Six mutants have been identified as having statistically significantly lower toxicity than the wild type strain ($p<0.05$, one-tailed Student's T-test), as shown in figure 4.4. Results are ordered numerically, first based on the locus tag of the gene and secondly based on the position within the gene at which transposons have inserted. Groups of transposon mutants that are within the same gene are outlined in black on the X axis labels.

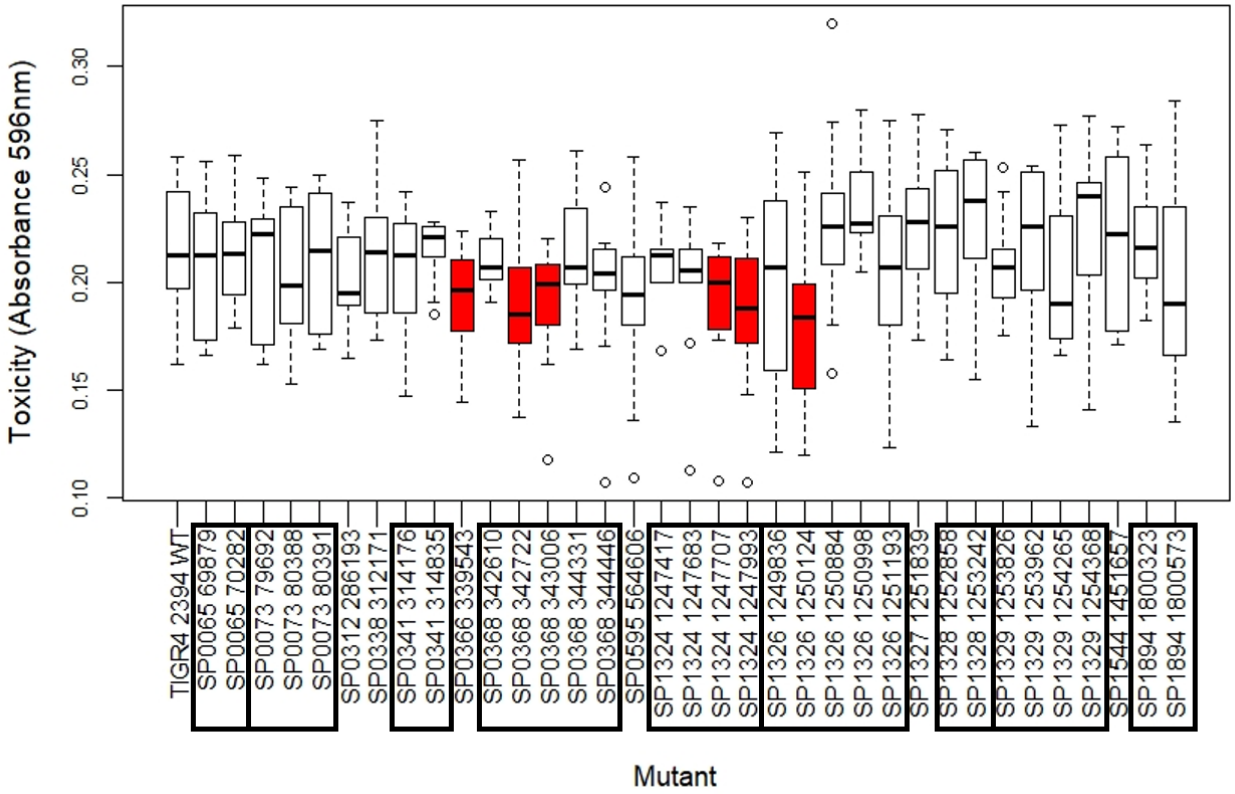


Figure 4.4: Average toxicity of each TIGR4 2394 mutant based on ten repeats of the haemolytic activity assay. Wild type TIGR4 2394 (first box) was used as a control, and as such all statistical tests for significance were done in comparison to this isolate. Data highlighted in red are mutants that are significantly less toxic than the wild type ($p<0.05$, one-tailed Student's T-test). Boxes represent the interquartile range, thick black lines within the boxes represent the median, whiskers represent the maximum and minimum values, with outlying data points represented as circles outside the whiskers. Black boxes around the X axis labels represent groups of mutants that are different transposon insertions within the same gene.

Results highlighted in red in figure 4.4 are mutants with a significant reduction in toxicity compared to the TIGR4 2394 wild type strain. Details of these mutants are listed in table 4.6 below.

Table 4.6: **Transposon mutants with reduced toxicity.** Mutants identified as having statistically significantly lower toxicity than the TIGR4 2394 wild type strain.

Gene locus tag in TIGR4 (location of Tn insertion)	Gene name in SPN23F	Gene function
SP0366 339543	aliA	putative extracellular oligopeptide-binding domain
SP0368 342722	SPN23F03400	cell wall surface anchored protein
SP0368 343006	SPN23F03400	cell wall surface anchored protein
SP1324 1247707	SPN23F12160	putative ROK family protein
SP1324 1247993	SPN23F12160	putative ROK family protein
SP1326 1250124	SPN23F12180	putative sialidase (neuraminidase)

Interestingly, where more than one mutant was provided with transposon insertions at different locations within the same gene, not all these mutants exhibited the same loss in toxicity. For example, the mutants with transposon insertions at positions 1247707 and 1247993 within the gene SP_1324 are two of four mutants with transposon insertions within this gene, and yet the other two mutants do not exhibit significantly reduced toxicity compared to the wild type. Both these insertions are within the first half of the gene after the start codon (see figure 4.5), and so the remaining two mutants could be producing a truncated form of the protein which may still retain some degree of normal function.

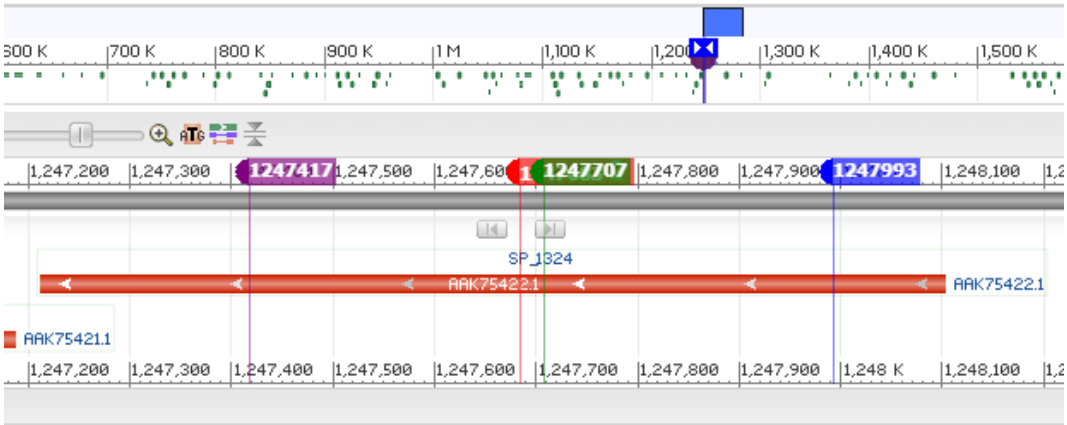


Figure 4.5. TIGR4 2394 gene SP1324, labelled with the positions of each transposon insertion. Transposon insertions at positions 1247707 and 1247993 are marked green and blue, respectively.The purple and red markers represent the two insertions which do not cause a loss in toxicity.

However, the insertions at positions 1247707 (low-toxic, green marker in figure 4.5) and 1247683 (toxic, red marker in figure 4.5) are in very close proximity, so it is unexpected that one would cause a significant reduction in toxicity while the other wouldn't; the transposon used to disrupt the gene would likely be disrupting the same structural feature of the resulting protein given the close proximity of these two loci, so it is unexpected that such a substantial difference in toxicity is observed between these two particular mutants. Furthermore, of five transposon insertions in the gene SP0368, only the second and third insertions (from the start codon) exhibit a reduction in toxicity (see figure 4.4).

It is noted that even after 10 repeats of the toxicity assay the variance of this dataset is high, as demonstrated by the lengths of the whiskers on the box and whisker plots in figure 4.4, therefore it is difficult to draw any firm conclusions from these results. Furthermore, the wild type TIGR4 2394 background strain in which these mutants were constructed, and which is used here as a control, is an unencapsulated pneumococcal strain which has lower toxicity than the encapsulated clinical strains used in the GWAS (absorbance at 596nm of 0.15-0.25 compared to 0.25-0.45 for the TIGR4 wild type), so variation between mutants and wild type is likely to be less pronounced. To determine if any differences in toxin production could be observed in the six mutants exhibiting a significant reduction in toxicity, a Western blot was conducted on cell pellet extractions using anti-pneumolysin antibody. The results of this are shown in figure 4.6 below.

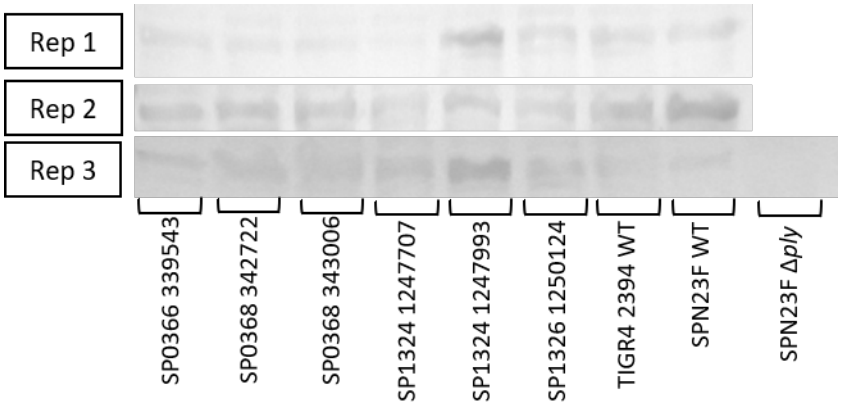


Figure 4.6. Anti-pneumolysin western blots. Cell pellet extractions from each of the TIGR4 2394 transposon mutants which showed a reduction in toxicity were used in a Western blot using anti-pneumolysin antibody, along with the TIGR4 2394 wild type, the SPN23F WT and, for the final repeat, a pneumolysin-deficient SPN23F mutant.

Here, mutant SP1324 1247993 appears to produce more pneumolysin than the two wild type controls; when quantified with ImageJ there is on average a 1.4-fold increase in pneumolysin production in this mutant in comparison to the TIGR4 2394 wild type control. Likewise there is on average a 2.5-fold decrease in pneumolysin production in the mutant SP1324 1247707 in comparison to the TIGR4 2394 wild type. This is an interesting result in that these two mutants contain transposon insertions at different positions within the same gene (SP1324 in strain TIGR4, SPN23F12160 in strain SPN23F). The western blot results for mutant SP1324 1247993 also contradict the results of the toxicity assay, in which this mutant was determined to be significantly less toxic than the wild type.

The fact that the unencapsulated TIGR4 2394 background strain is less toxic than the encapsulated TIGR4 strain could explain why it is difficult to distinguish differences in toxicity in this collection of mutants using the haemolytic activity assay. Western blotting may be more indicative of changes in toxicity in these mutants, however, as these results contradict those of the haemolytic activity assay, investigation of these mutants was not taken any further. It was decided a different approach was needed to screen the large number of GWAS hits.

4.2.3.2 Construction of whole gene knockout mutants in ATCC 700669 (SPN23F)

Given the variability and inconsistencies in the results from the transposon mutant collection in the background TIGR4 2394, it was thought a better approach would be to construct whole-gene deletion mutants in the high-toxic GWAS reference strain ATCC 700669 (SPN23F) in as many toxicity-associated genes as possible. A review of the literature (Ge & Xu. 2012; Chen et al., 2015) suggested that high-throughput mutant construction in *S. pneumoniae* was a relatively straightforward process in this naturally competent bacterium. Initially, it was planned to construct mutants in as many associated genes as possible from each of the three different GWAS methods, and to compare which method or combination of methods produced the most accurate results.

It was also thought that whole gene deletion was a better strategy than making SNP changes at the specific loci which were associated with toxicity. This was in part due to the sheer number of statistically significant loci identified, but predominantly because slight differences in toxicity which may be present in strains with a SNP compared to those without may not be visible using the

haemolytic activity assay, while whole gene deletion is more likely to have a pronounced effect on the phenotype.

To begin the mutant construction process, a Δ*ply* mutant deficient in the pneumolysin gene was constructed for use as a negative control in subsequent phenotypic assays. For in-depth description of the method used for mutant construction, see chapter 2, section 2.3.4. Briefly, overlap extension PCR was used, in which three DNA fragments are amplified; one homologous to the 1kb region upstream of the associated gene, one containing the antibiotic resistance cassette (in this case *ermAM*, amplified from pVA838) and one homologous to the 1kb region downstream of the associated gene. These fragments are designed to overlap each other, such that they can be stitched together in a separate PCR reaction to construct a linear insert around 2.75kb in length, with the antibiotic resistance cassette in place of the associated gene (see chapter 2, section 2.3.4.3 for details of the primer design process). This can then be transformed into *S. pneumoniae* ATCC 700669 (SPN23F) where it will recombine into the chromosome, deleting the associated gene. Figures 4.7 & 4.8 demonstrate the results of each PCR step to generate the *ply* 2.75kb insert.

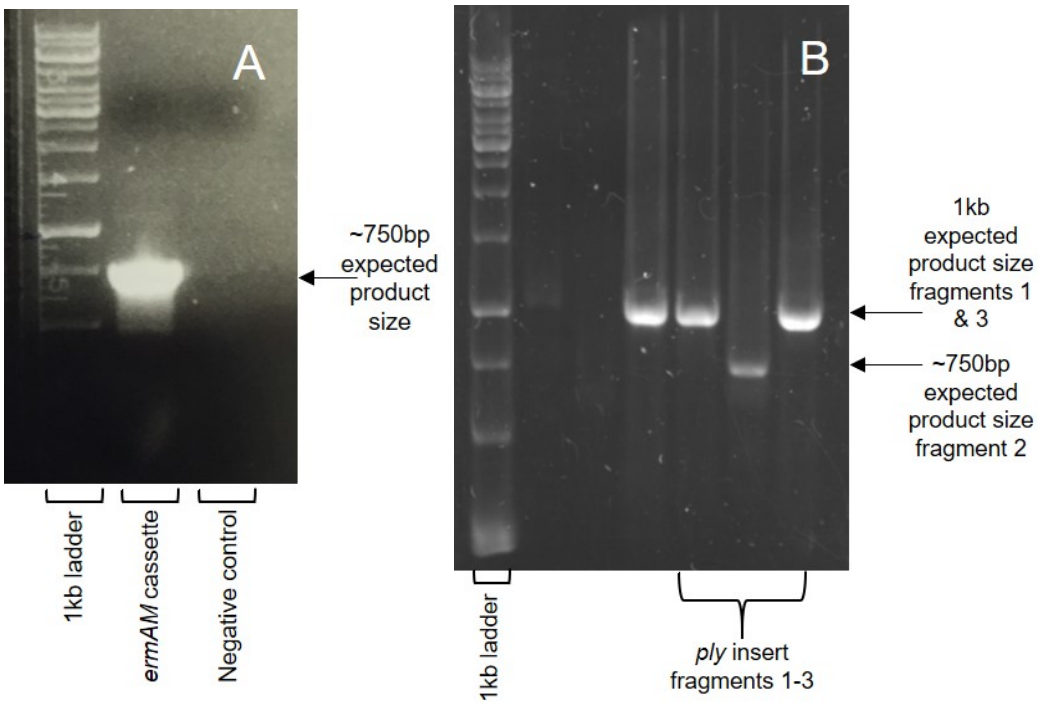


Figure 4.7. Initial PCR steps for mutant construction. A) Amplification of *ermAM* antibiotic resistance cassette from pVA838 for use as a template for amplification of fragment 2 for each insert. **B)** Amplification of fragments 1-3 for construction of the *ply* insert.

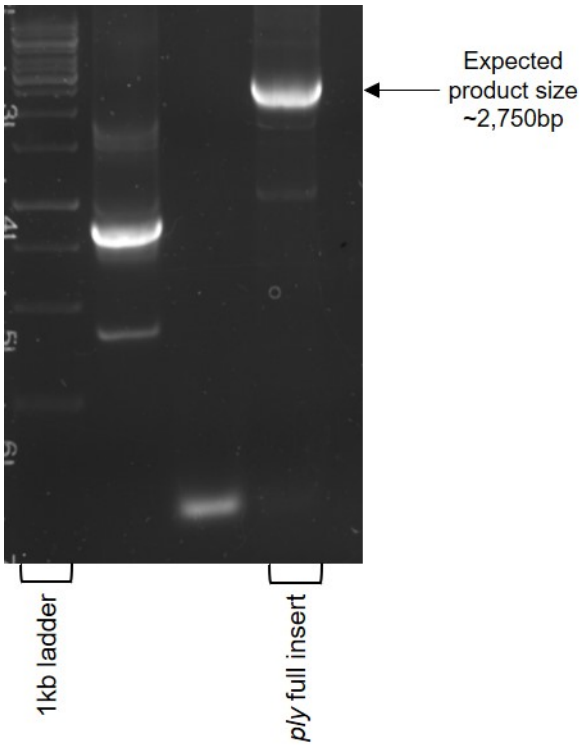


Figure 4.8. Stitch PCR for construction of *ply* 2.75kb insert. Fragments 1-3 were combined into one PCR reaction, in which the overlapping ends of each fragment were stitched together using the forward primer for fragment 1 and the reverse primer for fragment 3, creating a single linear insert around 2.75kb in length.

This 2.75kb insert was then transformed into *S. pneumoniae* strain ATCC 700669 (SPN23F). Successful transformants were verified using three PCR checks: one to a different *S. pneumoniae*-specific gene to verify the transformant isn't a contaminating organism (for all other mutants primers to the *ply* gene were used for this check, in this case primers were used to gene SPN23F13160 of strain ATCC 700669 [SPN23F]); one to the wild type gene that has been deleted; and one to the region 1kb upstream of the 2.75kb insert to within the *ermAM* cassette to verify both that the insert is there, and that it has inserted into the correct position on the chromosome. Figures 4.9, 4.10 & 4.11 below demonstrate these PCR checks for the Δ*ply* mutant.

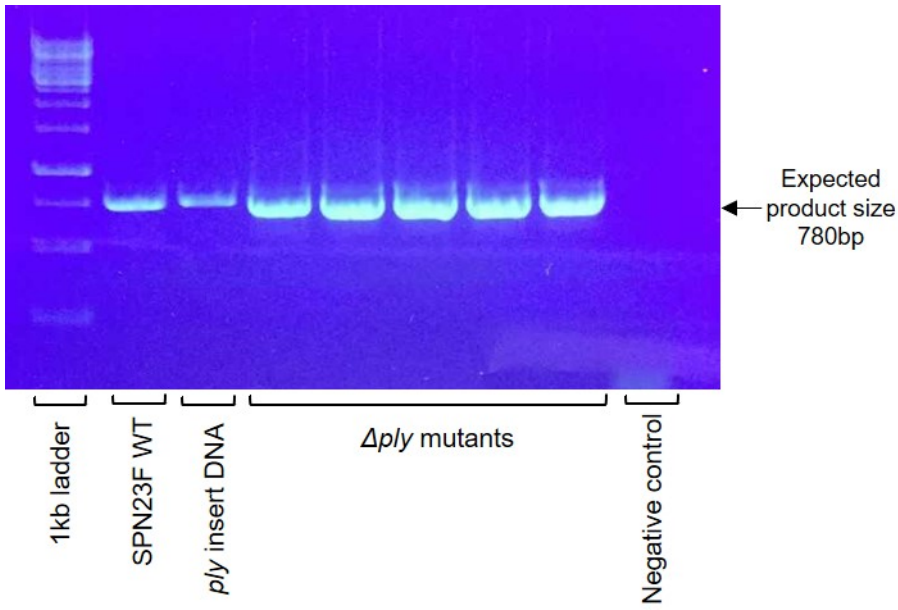


Figure 4.9. Verification that successful transformants are not contaminating organisms. PCR was conducted using primers to the SPN23F13160 gene of strain ATCC 700669 (SPN23F), to verify that successful transformants were *S. pneumoniae*.

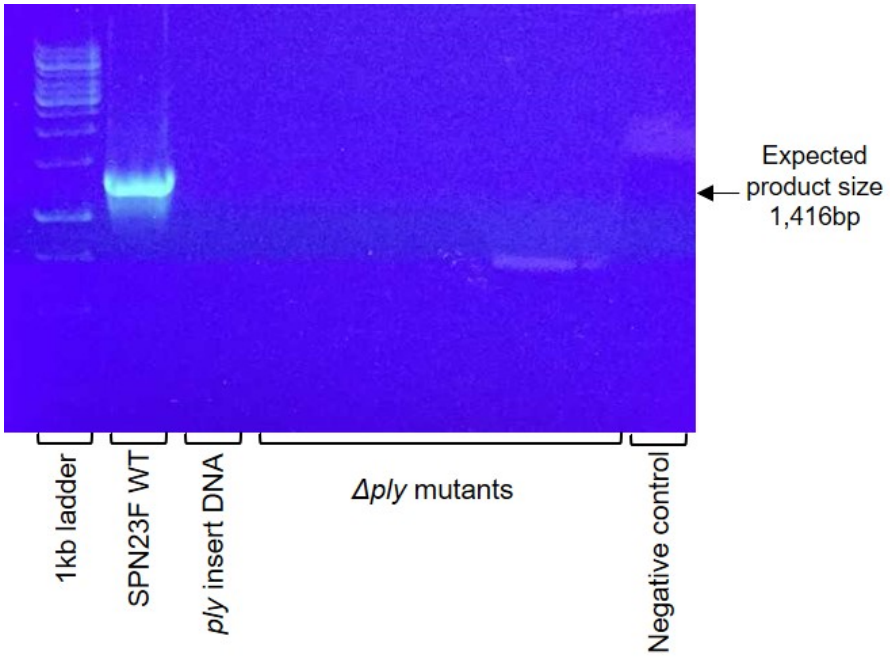


Figure 4.10. Verification that the Δ*ply* mutants do not contain the wild type gene. PCR was conducted using primers to the wild type *ply* gene. Results show amplification for the positive control only (SPN23F WT), as expected.

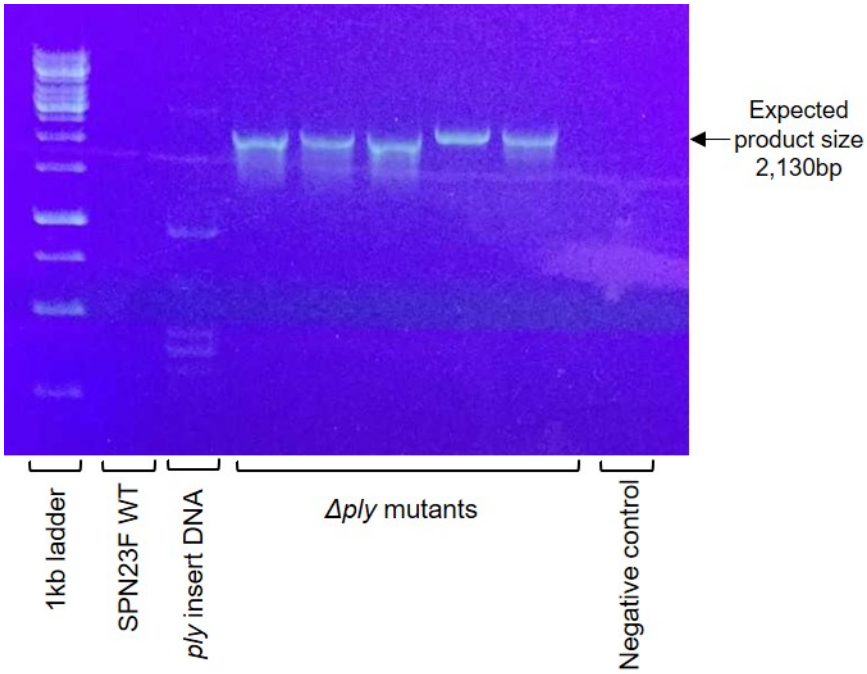


Figure 4.11. Mutant verification for Δ*ply* mutants. PCR was conducted using a forward primer placed 1kb upstream of the 2.75kb transformed insert, and a reverse primer placed within the middle of the *ermAM* cassette. Results show amplification for the successfully transformed mutants, but not for the SPN23F WT, or the *ply* insert DNA controls, as expected.

After verification of the successful construction of a Δ*ply* whole gene knockout, mutant construction was attempted for as many toxicity-associated genes from each of the GWAS methods as possible. Genes were prioritised for mutant construction on the following basis:

- Genes identified by more than one GWAS method (eg. *trcF*, *traG*, *priA* etc.)
- Genes from the group of TIGR4 2394 transposon mutants which were shown to have reduced toxicity (eg. *aliA*, SPN23F03400, SPN23F12160, & SPN23F12180)
- Genes which were thought to be likely to affect toxicity based on current knowledge of their function and correspondence with collaborators (eg. *nanA*)
- All other genes – inserts were constructed and transformed for all remaining genes in the order of first to last genes in the results tables

A number of signature-tagged mutagenesis screens could also have been used at this stage to aid the process of defining genes most likely to be involved in virulence. Examples of these include Polissi et al. (1998), Lau et al. (2001), Hava

& Camilli (2002) and Obert et al. (2006). Hava & Camilli, for example, identified a number of tissue-specific virulence factors in a serotype 4 pneumococcal strain, some of which were also identified as significant associations in the GWAS analyses presented here (eg. *gidB*, *ntpK* and *msmF*).

To date, 2.75kb inserts have been constructed for 59 associated genes. Figures 4.12 & 4.13 give examples of successful insert construction for the first four of these.

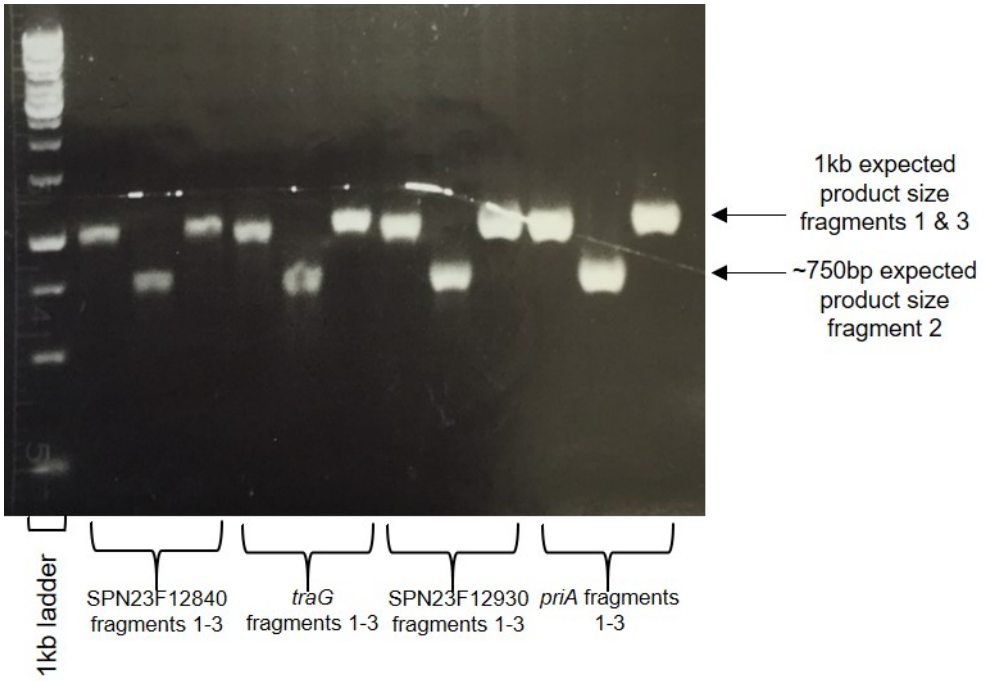


Figure 4.12. Fragment amplification for the first four inserts. Successful amplification of fragments 1-3 for the genes SPN23F12840, *traG*, SPN23F12930 and *priA*, all of which were associated with toxicity by both the SNP-based GWAS method and SEER.

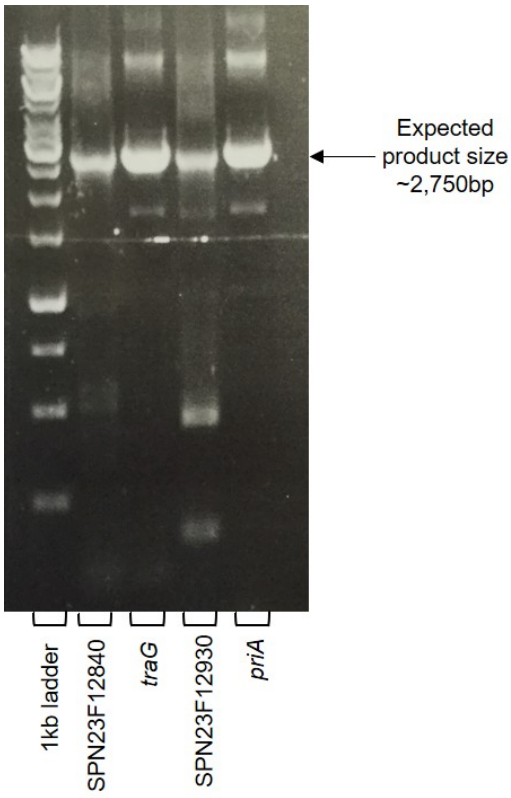


Figure 4.13. Stitch PCR for the first four inserts. 2.75kb inserts were stitched together using the forward primer for fragment 1 and the reverse primer for fragment 3.

Transformations were attempted for each of the 59 inserts at least once, with success for 9 associated genes as listed below.

- *nanA* – probable N-acetylneuraminate lyase
- *priA* – putative primosomal protein N
- SPN23F12470 – putative DNA helicase II, UvrD-like protein
- SPN23F01610 – putative ABC transporter system permease protein
- SPN23F01620 – putative membrane protein
- SPN23F02870 – putative glycosylhydrolase
- SPN23F15290 – autolysin
- *trcF* – transcription repair coupling factor
- *gidB* – methyltransferase GidB

Difficulties in the mutant construction process were thought to be primarily due to inefficiency of the background strain in transformations, this is discussed further below and variations that were made to the protocol to try and increase efficiency are presented in table 4.7.

Figures 4.14 & 4.15 below illustrate the mutant verification process using *priA* as an example, which was the only one of the four inserts shown in figure 4.13 to transform successfully.

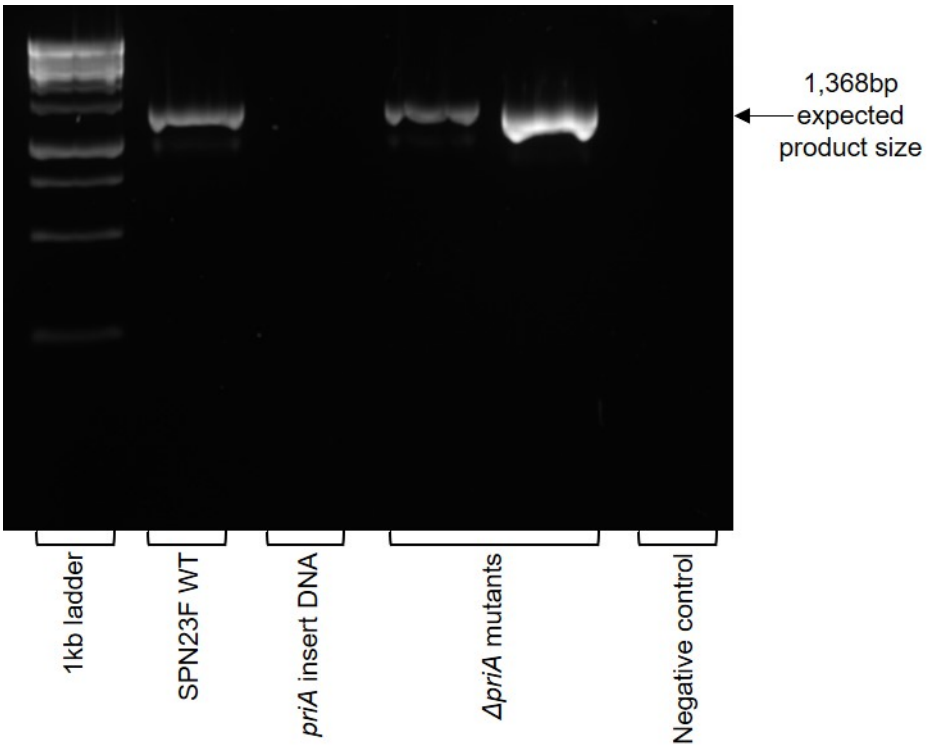


Figure 4.14. Verification that Δ*priA* mutants are not contaminating organisms. Two Δ*priA* transformants were obtained, which are verified here to be *S. pneumoniae* and not a contaminating organism using primers to the wild type *ply* gene.

Whilst it was initially thought that use of the wild type *ply* primers would demonstrate that transformants were colonies of *Streptococcus pneumoniae*, it was later established that a highly similar gene is present in the genome of *S. mitis*, so another gene should have been chosen for this step.

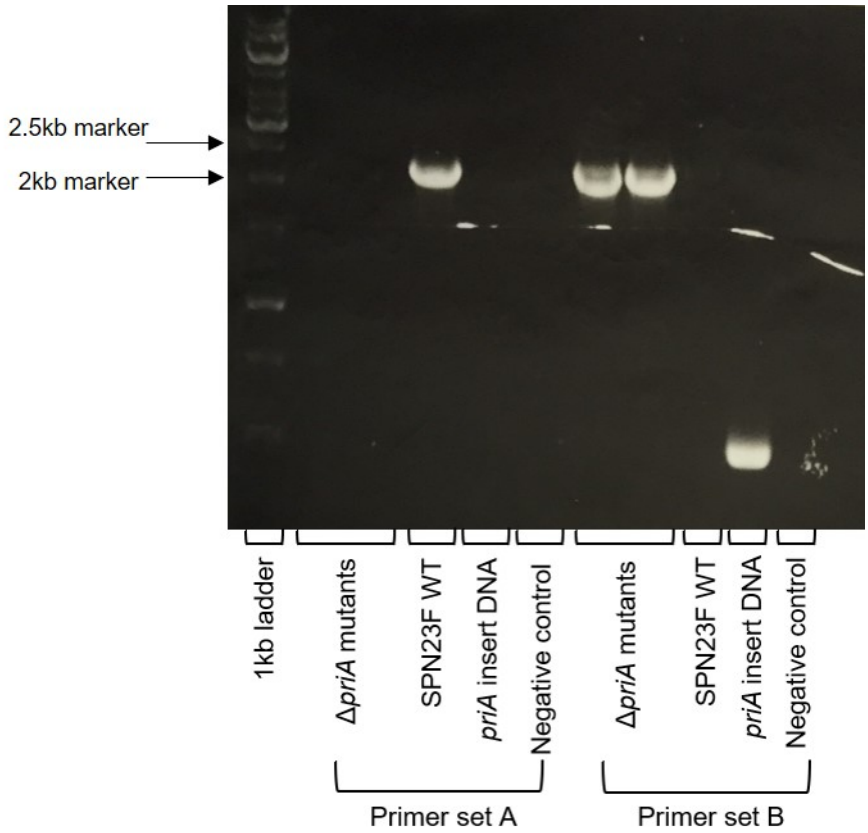


Figure 4.15. Verification of successful gene deletion in Δ*priA* mutants. Primer set A - primers to detect the wild type *priA* allele (expected product size 2397bp), with expected amplification for the SPN23F WT DNA, but not for the Δ*priA* mutants or the *priA* insert DNA. **Primer set B -** primers to amplify the region from 1kb upstream of the transformed 2.75kb insert to within the centre of the *ermAM* resistance cassette (expected product size 2430bp), with expected amplification for the Δ*priA* mutants, but not for the SPN23F WT DNA or the *priA* insert DNA.

Each of the 9 whole gene deletion mutants were subsequently tested for toxicity using the haemolytic activity assay, using wild type ATCC 700669 (SPN23F) and the SPN23F Δ*ply* mutant as positive and negative controls, respectively. Results for each mutant are shown below in figure 4.16.

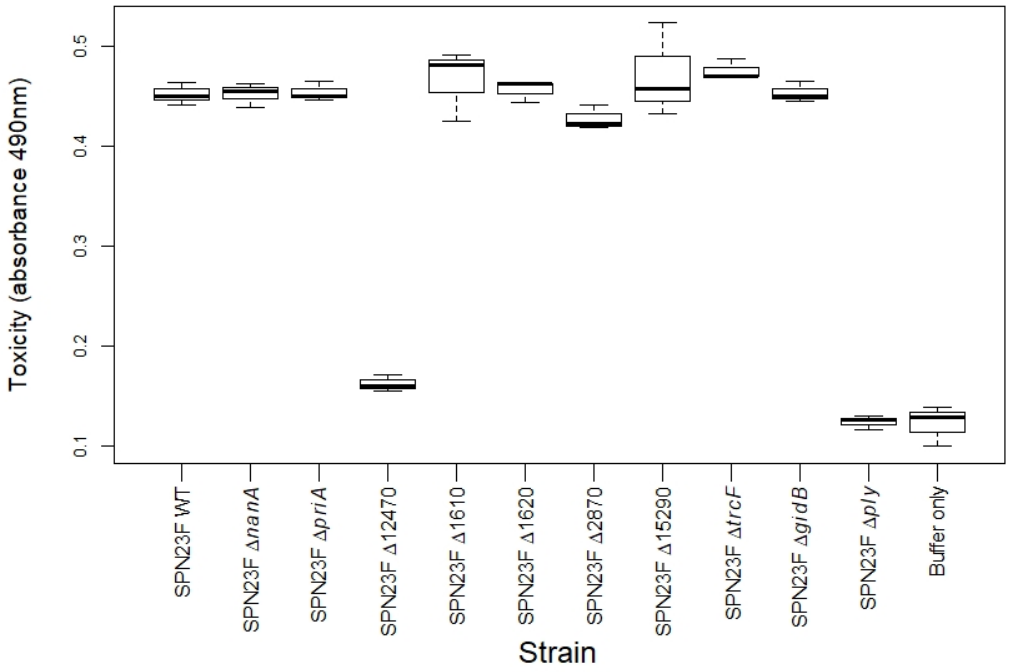


Figure 4.16. Average toxicity of SPN23F whole gene deletion mutants. Toxicity of each SPN23F whole gene deletion mutant was tested using the haemolytic activity assay, results shown are the averages of at least three biological repeats. Controls used were the SPN23F wild type strain (first box), the pneumolysin-deficient mutant and buffer only (last two boxes).

Of the 9 mutants constructed here, one mutant deficient in the gene SPN23F12470 (identified by SEER) shows a significant reduction in toxicity in comparison to the isogenic wild type similar to the reduction in toxicity observed for the Δ*ply* mutant. Detailed characterisation of the role this gene plays in relation to toxicity is presented in chapter 5.

Of the genes which were identified by more than one GWAS method (SPN23F12840, *traG*, SPN23F12930, *priA*, *trcF*, *pbpX*, *mraY*, SPN23F03400), mutants were successfully created in the *priA* and *trcF* genes, despite several attempts to transform inserts for the other genes. After inconsistencies in transformation efficiency were observed, several variations to the protocol were used in separate experiments to attempt to determine why the process was not working consistently. These are listed in table 4.7 below.

Table 4.7. Variations to transformation protocol for ATCC 700669 (SPN23F). Several variations to the protocol were introduced to try to improve transformation efficiency. Results recorded are based on observation of the number of colonies which grew on transformation plates over numerous repeated experiments.

Change to protocol	Result (observation only)
Introduction of a positive control using original *ply* insert	Inconsistent results - transformations were inconsistent for both this control insert and genes of interest
Varied amount of competence stimulating peptide 2 (CSP-2) from 2-20ng/ul	No change in transformation efficiency
Varied amount of DNA added to competent cells	No change in transformation efficiency
Varied starting density of cells using up to 6 serial 10-fold dilutions of inoculum for overnight culture	No change in transformation efficiency
Varied recovery time from 30 minutes to 2 hours	No change in transformation efficiency

After several attempts at transforming each insert, it was established that transformation of the ATCC 700669 (SPN23F) wild type strain is inefficient. This was corroborated through communication with Tim Mitchell at the University of Birmingham and Nick Croucher at the Sanger Institute, who confirmed that this background strain is less efficient in the transformation process than other pneumococcal strains. For this reason, it is difficult to establish whether many of the inserts which were used in attempted transformations did not produce viable mutants because they were disrupting essential genes, or whether it was due to the inefficiency of the background strain.

4.3 Discussion

Genome-wide association studies (GWAS) are increasingly being applied to studies of bacterial phenotypes thanks to advances in sequencing technology and a subsequent increase in the number of complete bacterial genomes which have become available. Since the first study which applied GWAS to bacterial genomes (Sheppard et al., 2013), the approach has been gathering momentum, with several more studies being published which use GWAS to identify genetic associations with a phenotype (Farhat et al., 2013; Alam et al., 2014; Laabei et al., 2014; Chen & Shapiro, 2015; Schubert et al., 2019).

Due to the high number of loci identified as significantly associated with toxicity from all three methods used here, and the lack of a library of mutants such as the Nebraska Transposon Mutant Library which was used for functional verification of

GWAS in *Staphylococcus aureus* (Laabei et al., 2014; 2015), it was thought that comparison of results between GWAS methods would a be useful process to aid prioritisation of genes for further investigation. It was hypothesised that true positive loci would consistently appear in the results of different GWAS methods despite variations in the approach used, and that it would in turn be easier to both identify likely false positive results and prioritise true positive genes for functional verification.

Comparing the results of different methods was also done to try to reduce the likelihood of wrongly excluding loci which directly affect the phenotype; i.e. the likelihood of obtaining false negative results. Furthermore, following the rapid increase in the number of studies using GWAS to investigate bacterial phenotypes, and the resulting diversity of methods used, it was of interest to compare the results of different methods to determine whether one, or a combination of methods, gives a more accurate representation of true genetic associations with a phenotype.

4.3.1 SNP-based GWAS

The SNP-based GWAS approach used to gather the first set of data (table 4.1) is based on identifying statistically significant correlations between the presence of individual SNPs at loci throughout the genome and high or low levels of toxicity using Pearson's product-moment correlation. A potential limitation of this approach when used in isolation is that individual SNPs that cause a minor change to the resulting protein may not correlate very highly with changes in toxicity, while the whole gene containing that SNP may be functionally important in regulating, producing or releasing pneumolysin; this will appear as a false negative in the results of our GWAS analysis. This is the traditional method used in GWAS to identify genetic associations with the phenotype, but combining this method with the other two approaches used here was thought to minimise the potential for exclusion of false positive results. Likewise, since SNPs close together in the genome may be inherited together during recombination, while only one may in fact be associated with the toxic phenotype, this approach may incorrectly identify many loci as being statistically associated with the phenotype when they are not, and indeed this approach can be prone to producing high numbers of false positive results. Again, combining this method with the other two approaches was thought to minimise the potential for false associations with the phenotype.

Highly conserved loci in which there are very few SNP changes will also be excluded from the analysis in the initial data filtration step, and therefore could be wrongly considered a negative result; this is an artefact of the initial data filtration step aimed at reducing the number of statistical comparisons that need to be made to reduce the computational intensity of the analysis. A good example of this is the pneumolysin gene *ply*, which is a highly conserved locus that has a direct effect on toxicity in *S. pneumoniae*. Due to the lack of SNP changes within this gene this region was excluded from the analysis in the initial data filtration step based on the minor allele frequency threshold of 5%, and therefore all loci within this gene are discounted.

The SNP-based approach does not correct for population structure, and so causative variants that are present only within small subsets of the population may confound the results, reducing the statistical power of the analysis and increasing the number of false positive results. To some degree, the inverse may also be true, whereby causative variants are missed precisely because they are only present in a small subset of the population. Controlling for population structure in GWAS relies on a trade-off between these two situations; a reduction in false positive associations can lead to an increase in false negative associations and vice versa. However, all strains used in this collection are taken from the same clonal complex, PMEN-1, and *S. pneumoniae* is a highly recombinogenic bacterial species; as such it was thought this would reduce the confounding effects of population structure on the SNP-based GWAS, so it was not corrected for in this method.

Application of the Bonferroni correction to this dataset resulted in no loci crossing the threshold of statistical significance as the cut off was too stringent. This reduced the power of the analysis effectively to zero, however, visualisation of the results in a Manhattan plot without application of the Bonferroni correction clearly showed peaks of significant loci clustered at several points across the genome. It was decided that a permutation test would be a better alternative to correct for multiple tests in this case, in which the dataset is scrambled up to a thousand times to determine the likelihood of a locus being associated with the phenotype by chance alone. P-values are then re-calculated to give empirically derived results in which significance is determined not to be an artefact of chance. Figure 4.3 shown in results section 4.2.2.1 shows the result of the SNP-based analysis after permutation, with a significance threshold set at p=0.05, in which peaks of significant loci can be clearly seen. While permutation is a useful

method of reducing the false positive error rate in an analysis with such a high number of statistical comparisons, functional verification is still essential to eliminate false positive results.

4.3.2 SEER

SEER (sequence element enrichment analysis), the method developed by John Lees, uses logistic and linear regression to look for associations between k-mers and phenotypic variation; k-mers are sections of DNA of length 'k' (usually 7-35 nucleotides) which can encompass numerous types of genetic variation across a collection of genomes including insertion and deletion of genetic elements, changes in promoter architecture, and recombinations, to name a few examples (Lees et al., 2016). This increases the statistical power of the analysis and produces more biologically relevant results, because SNPs, although a common effector of phenotypic change, are not the only genetic elements responsible for a change in bacterial phenotypes. Use of k-mers therefore widens the scope of the analysis and reduces the extent to which genetic loci are discarded as non-significant associations when they are in fact causative of a change in the phenotype.

A second major difference between the SNP-based approach and SEER is that SEER corrects for clonal population structure within the collection of pneumococcal genomes using metric multi-dimensional scaling (mMDS). Lees et al. (2016) describe the metric multi-dimensional scaling method as being analogous to use of principal components of the SNP matrix in human genetics to control for divergent ancestry. mMDS was preferentially used here because it does not rely on core gene alignment or SNP calling, but can be directly applied to the k-mer counting result to accurately estimate bacterial population structure. The benefit of controlling for population stratification in this way is that it reduces the rate of false positive associations by excluding from the results those variants that are causative of a change in the phenotype in only a small subset of the population, however, the threshold at which these associations are excluded from the results could cause an increase in false negative associations if set too high.

Unlike the SNP-based approach, SEER utilises the Bonferroni adjustment to correct for multiple testing but does not correct for every test conducted. A conservative adjustment is made to the threshold $p<0.05$, which is corrected to 1×10^{-8} to prevent a high number of false positive associations while not over-correcting based solely on the number of tests conducted. Application of the

Bonferroni correction to the SNP-based results demonstrated that the cut-off for significance is too stringent using the traditional Bonferroni calculation, as it made every result from the analysis non-significant, while here a conservative adjustment is made so that power is not lost from the analysis but the number of false positive results is reduced.

Table 4.2 shows that SEER produced a smaller list of genes which were statistically associated with toxicity, four of which were also identified by the SNP-based approach. Mutant construction was successful for one of these - the *priA* gene - which was then verified to be false positive. Mutant construction for the remaining three genes was unsuccessful, so functional verification of these results could not be completed. However, SEER is the only method from which a true positive association with toxicity has been identified at the time of writing. The SPN23F12470 gene identified by SEER showed a significant reduction in toxicity in comparison to the isogenic wild type strain when tested for haemolytic activity (figure 4.6), and this gene was not identified using either of the other methods. In depth characterisation of the relation of this gene to toxicity is presented in chapter 5.

4.3.3 Bugwas

The third method used to contribute to the dataset was Bugwas, which uses linear mixed models (LMMs) to look for associations between genotype and phenotype. Bugwas makes use of both SNPs and k-mers as the units of genetic variation by conducting separate analyses on both, the results of which are split into two groups for comparison (tables 4.3 and 4.4). Bugwas further divides the SNP results into bi-allelic and tri-/tetra-allelic loci and the k-mer results into fixed-length and variable-length k-mers. The original SNP-based approach does not distinguish between bi-, tri- or tetra-allelic loci, and SEER uses only variable length k-mers. Similarly to SEER, Bugwas controls for clonal population structure within the collection of genomes, and applies the Bonferroni adjustment to correct for multiple testing.

The use of SNPs, fixed length k-mers and variable length k-mers allows an interesting comparison between three different sets of results from the same analysis. The number of SNPs statistically associated with the toxic phenotype is substantially larger than the number of k-mers statistically associated with the phenotype, even with the results of the fixed and variable length k-mers combined. There is crossover between the two datasets at the *pbpX* and *mraY*

genes, suggesting a strong association with the phenotype which is supported by the simultaneous identification of *pbpX* and *mraY* in the SNP-based method. Unfortunately, mutant construction was unsuccessful for both genes, so functional verification of these results has not been possible.

As with SEER, Bugwas accounts for population structure in the GWAS analysis, but specifically allows for identification of groups of lineage-specific variants that collectively show a strong association with the phenotype. Bugwas uses a slightly different approach to SEER to do this, using principal components to identify lineage-level differences in association with the phenotype. However, it does not eliminate the need for functional verification, as controlling for population structure in this way requires a trade-off between detecting causative groups of variants within different lineages of the population that show a collective association with the phenotype and causative variants present in only a small subset of the population which may only serve to confound the results.

4.3.4 treeWAS

A fourth method was attempted called treeWAS, originally developed by Caitlin Collins and Xavier Didelot at Imperial College London (Collins & Didelot, 2018). The aim of this approach is to account for population stratification, much like Bugwas and SEER, but by using the genomic dataset to build a phylogenetic tree from which population structure is inferred prior to the GWAS analysis. However, this method produced no significant hits because use of the traditional, highly stringent Bonferroni correction is built into the program, and it cannot be excluded from the analysis. As with the SNP-based GWAS, application of the Bonferroni correction meant the threshold for significance was too high for any of the results to be significant. For this reason, the treeWAS results have not been included here.

4.3.5 Collective GWAS results

It is particularly interesting that no one gene has been identified as statistically significantly associated with toxicity in all three of the GWAS methods used. Additionally, the pneumolysin gene has not been identified as an effector of toxicity using this method. This is because the gene is highly conserved within the PMEN-1 clonal complex, and as such the number of SNPs present within this gene were too low to cross the minor allele frequency threshold of 3%, which

requires there to be a minimum of 3% SNP changes present at each locus in the clinical collection of strains in order to be included in the subsequent GWAS.

While there is some crossover of genes that are identified in two of the three methods, there are relatively few genes which are identified by more than one method. In total, 8 genes were identified by more than one method out of a combined total of 181 genes and intergenic regions across the three approaches.In some cases, several genes in close proximity to each other have been individually identified by one GWAS method, while other genes within the same region have been identified in another method. For example, genes within the region from *agaS* to SPN23F00850 were individually identified by both the SNP-based GWAS method and Bugwas. Genes from *agaD* to *clpL* are likewise identified by these two methods. A large cluster of genes on the ICE*Sp*23FST81 mobile genetic element are identified by both the SNP-based method and SEER, and a small number of genes close to the *priA* locus are identified by all three methods.

This implies that all three methods combined may be useful for identification of specific clusters of loci on the chromosome where the search for toxicity-affecting loci should be focussed, however, no one method or combination of methods seems at this point to be preferable for identification of individual true positive loci which directly affect the phenotype. Data from signature-tagged mutagenesis screens could also be added to these results to better identify true positive associations with toxicity (eg. Polissi et al., 1998; Lau et al., 2001; Hava & Camilli, 2002; Obert et al., 2006). Furthermore, without successfully constructed mutants in and around these clusters of loci, their potential role in toxicity cannot be determined.

4.3.6 Functional verification

Tentatively, four genes have been identified as causing a significant loss in toxicity in the unencapsulated background strain TIGR4 2394 in comparison to the isogenic wild type strain. These genes include a putative extracellular oligopeptide binding domain (*aliA*), a cell wall surface anchored protein (SPN23F03400), a putative ROK family protein (SPN23F12160) and a putative sialidase otherwise known as neuraminidase (SPN23F12180).

AliA is an oligopeptide-binding lipoprotein (Alloing et al., 1994) with high homology to the oligopeptide permease ABC transporter AmiA (Claverys et al.,

2000). The *aliA* gene neighbours the capsular polysaccharide biosynthetic pathway in *S. pneumoniae* type 19F, although it is not involved in this pathway (Morona et al., 1997). The SPN23F03400 locus is annotated as a cell wall surface anchored protein, and immediately neighbours the *aliA* gene in strain ATCC 700669 (SPN23F). However, the specific role of this gene appears to be uncharacterised, with only one reference to it in the literature which sequenced genes around the *cps* (capsular biosynthetic pathway) locus in *S. pneumoniae* to study the evolution of multi-drug resistance in serotype 6D (Ko et al., 2012).

ROK family proteins (Repressor, ORF, Kinase) include transcriptional repressors, sugar kinases and uncharacterised open reading frames (Titgemeyer et al., 1994). Neuraminidase, encoded by the gene SPN23F12180 in strain ATCC 700669 (SPN23F), is a known virulence factor of *S. pneumoniae* which cleaves terminal sugars from glycoconjugates (Kadioglu et al., 2008), and has been shown to be involved in invasion and inflammation of host tissues (Berry et al., 1989; Xu et al., 2011).

However, firm conclusions cannot be drawn regarding the role of these four genes in toxicity for the following reasons: firstly, the background strain used is an unencapsulated, low-toxic isolate, and as such variations in toxicity are more difficult to distinguish using the haemolytic activity assay; secondly, the variation in the toxicity data for these four genes is very high even after 10 repeats, so statistical significance may not reflect a true biological difference in toxicity in this case; and finally, in some cases transposon mutagenesis at different positions within these four genes does not result in a reduction in toxicity based on the results of the haemolytic activity assay. Whole gene deletion of these genes in the high-toxic reference strain ATCC 700669 (SPN23F) would provide clearer evidence as to whether these genes are causative of a change in the toxic phenotype. Where present, genes would also need to be deleted in other pneumococcal strains to determine whether their effect on toxicity, if any, is universal between strains.

Through the process of whole gene deletion in strain ATCC 700669 (SPN23F) eight genes have been identified as false positive associations, including the *priA* and *trcF* genes which were each highlighted by two GWAS methods. It is interesting that deletion of the SPN23F15290 gene, which is annotated as an autolysin, does not result in a significant loss of toxicity. Autolysis is thought to be key in releasing the pneumolysin toxin into the extracellular environment, as the

toxin does not contain an N-terminal secretory signal sequence, and therefore cannot be actively exported out of the cell. A possible explanation for the sustained level of toxicity in this mutant could be that since it is not interfering with pneumolysin production, or it's exportation to the cell wall where it is known to be located prior to autolysis of the cell, the toxin is still able to exert its effect on red blood cells during the hour-long incubation period through physical contact between bacterial cells and erythrocytes.

One gene – SPN23F12470 – has been identified as a true positive association with the toxic phenotype, and this was identified using SEER. This gene encodes a UvrD-like protein containing a helicase domain and two tandem repeat nuclease domains at the N-terminus. Detailed characterisation of how this gene affects toxicity in *S. pneumoniae* is presented in chapter 5.

4.3.7 Overview

Use of the GWAS method is the first step towards understanding the genetic basis of phenotypic variation. However, it is only useful for initial identification of causative genetic loci and does not reveal information about individual gene function in relation to a phenotype. If we are to work towards an improved ability to understand and interpret genome sequencing data in the context of disease, we must first understand the role of individual genes in pathogenesis.

GWAS is a top-down approach to achieving this goal, and by no means gives a complete picture of the genetic contributors to a phenotype as the method is prone to producing high numbers of both false positive and false negative results. Nonetheless, it is a useful starting point, and while the number of true positive associations identified thus far is extremely low, it has led to the identification of an interesting gene which would otherwise have remained unknown in its relation to toxicity in *S. pneumoniae*. Short of making a knockout mutant in every single gene within the pneumococcal genome and assaying each of these for their effect on toxicity, the role of the SPN23F12470 gene may otherwise have gone unnoticed. While it has its limitations, GWAS has therefore proven a useful tool in identifying a novel genetic contributor to toxicity in the pneumococcus.

4.4 Conclusions

GWAS has been used as a starting point in this project to identify genetic loci which have the highest probability of being associated with the toxic phenotype in *S. pneumoniae*. Comparison of the results from three different methods was

thought to be useful in better identifying true positive results without having to construct knockout mutants in every gene within the genome. However, the results presented here show that crossover of findings between methods was relatively low and due to the difficulties encountered in constructing mutants in the ATCC 700669 (SPN23F) background, functional verification of the results has been restricted to a relatively small number of genes.

Further, in-depth investigation of each of these genes is required to eliminate loci that have been falsely associated with this phenotype, and to identify loci which are directly associated with toxicity. From this it will be possible to build a more complete picture of the complex genetic network which is involved in regulation, production and release of the pneumolysin toxin.

The next steps for this project would be to continue construction of whole gene deletion mutants in the reference strain ATCC 700669 (SPN23F). Construction of a mutant in every associated gene from the GWAS analyses would allow verification of direct and indirect associations to be completed, from which firm conclusions could then be drawn regarding which GWAS method, or combination of methods, produced the highest number of true positive associations with toxicity. More importantly, completion of the mutant construction process would allow identification of true positive, novel associations with the toxic phenotype.

Short term, work such as this could lead to identification of novel drug targets for treatment of pneumococcal disease, but ultimately, research in the wider field is looking to better interpret the vast amount of genome sequencing information which is being gathered from an increasing number of bacterial species and strains. This information could prove useful in the long term if genome sequencing is to be used as a diagnostic tool for patients suffering from pneumococcal disease, to identify the specific strain causing infection and its corresponding pathogenicity and antibiotic resistance profile.

To achieve this, a thorough understanding of the genetic contributors to virulent phenotypes and antibiotic resistance profiles will be required, along with an ability to correctly interpret this information such that patients receive the correct treatment as efficiently as possible. There is a long way to go in improving our understanding of the genetic basis for bacterial virulence before such a system can be developed, and improvements in technology, the speed of sequencing, and the cost of such technology will also be required if this system is to become widely used in healthcare in the future.

Central to all of this will be the fundamental understanding of the genetic basis for bacterial virulence. The work presented here demonstrates how a statistical approach can be successfully applied to identify a novel genetic contributor to virulence in *S. pneumoniae* and has led to the identification of the novel effector of toxicity SPN23F12470.

5. Characterisation of a toxicity-deficient ΔSPN23F12470 mutant in *Streptococcus pneumoniae*

5.1 Introduction

Streptococcus pneumoniae produces only one toxin, pneumolysin (Ply), which is primarily responsible for the lysis of host cells during infection. Whilst the *ply* gene is the major contributor to toxicity in this species, it is demonstrated here that other genetic factors make a significant contribution to production of this toxin. In an era where whole genome sequences are becoming readily available to researchers, work such as this contributes to a better understanding of individual gene function in relation to a specific phenotype. Genome sequencing information by itself is useful in studying the evolution and epidemiology of a bacterial species, but if we are to start interpreting genome sequencing information or developing alternative treatments for disease, it is important to fully understand as much about the genetic contributors to pathogenesis as possible.

The work presented here follows on from the GWAS analyses discussed previously in chapter 4, which identified 181 genes and intergenic regions that are statistically significantly associated with the toxic phenotype in *S. pneumoniae* clonal complex PMEN-1. Through the process of functional verification of these GWAS results a whole gene deletion mutant was constructed in the gene SPN23F12470 in the background strain ATCC 700669 (SPN23F) which is shown here to have a deficiency in toxicity in comparison to the isogenic wild type.

SPN23F12470 is annotated as a putative DNA helicase II, homologous to the *uvrD* gene in other bacterial species such as *E. coli*. To determine how loss of function of this gene results in a loss of toxicity, work is presented here to elucidate the contribution of this gene and resulting protein product to pneumolysin production in the pneumococcus.

5.2 Results

5.2.1 Identification of the SPN23F12470 gene as a novel toxicity-associated locus in *S. pneumoniae*

Through the GWAS method SEER, described in the previous chapter, SNPs in the SPN23F12470 gene were statistically associated with toxicity in *S. pneumoniae*. A comparison of toxicity data for the original clinical strains with and without a SNP in the SPN23F12470 gene shows that strains with the polymorphic SPN23F12470 gene have on average a slightly higher toxicity than those with the reference SPN23F12470 gene, as shown in figure 5.1. A T-test between the two groups shows that this difference is significant (p=0.0018).

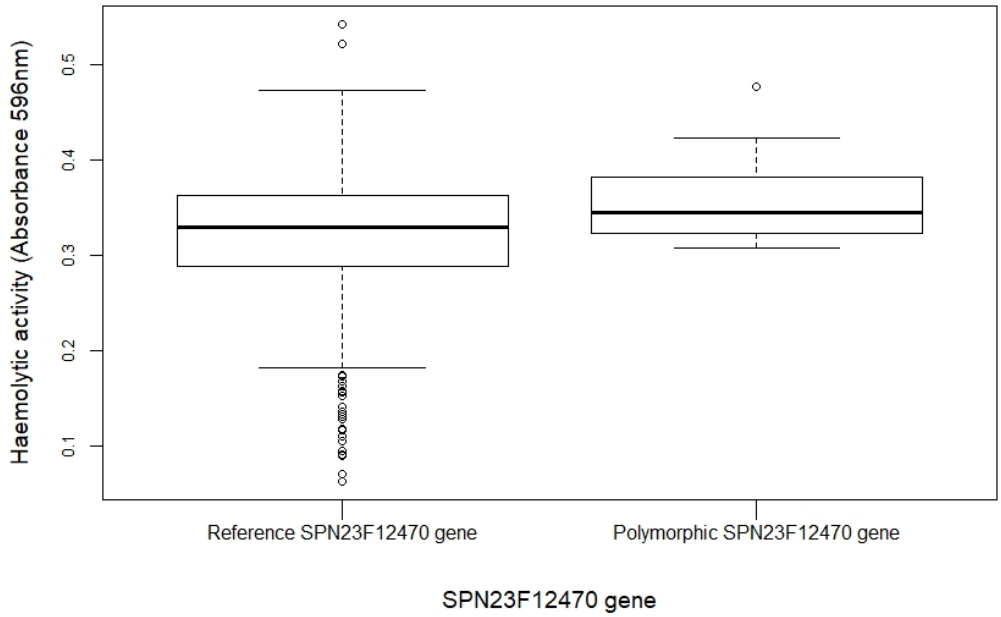


Figure 5.1. Comparison of toxicity of clinical isolates with and without SNPs in the SPN23F12470 gene. A significant difference is observed (p=0.0018) between clinical isolates with (n=6) and without (n=149) a SNP change within the SPN23F12470 gene, with higher toxicity observed on average in strains with a polymorphic SPN23F12470 gene.

5.2.1.1 SPN23F12470 encodes a UvrD-like helicase with two tandem repeat nuclease domains

The SPN23F12470 gene is annotated as a 'putative DNA helicase II' in strain ATCC 700669 (SPN23F). Upon discussion with Mark Dillingham (Dept. of Biochemistry, University of Bristol), it was ascertained that the gene encodes a

protein with three domains; a UvrD-like domain encoding a helicase motor for unwinding of DNA, and two tandem repeat domains encoding nucleases similar to those found in restriction enzymes and Cas4 proteins (see figure 5.2 below). Both tandem repeat domains contain an iron sulfur cluster region and are thought to be active nucleases, although this has not been confirmed. It was noted that this appears to be a highly unusual protein, when a BLAST is conducted to look for homologous genes in other species results show 98%-100% identical homologues in only two other strains of *Streptococcus pneumoniae*, two strains of *Streptococcus mitis* and two other strains of *Streptococcus* annotated as *Streptococcus sp.* JS71 and *Streptococcus sp.* A12. No other homology is identified.

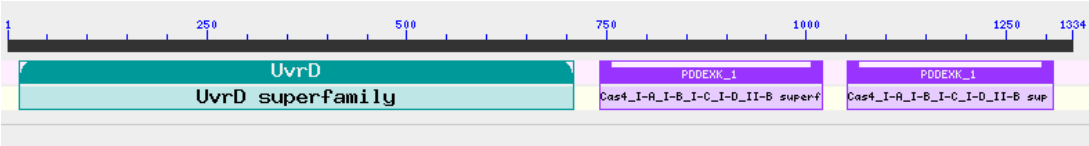


Figure 5.2. BLAST results for SPN23F12470. Results of protein BLAST for the SPN23F12470 gene in ATCC 700669 (SPN23F), showing the UvrD helicase domain and the two downstream tandem repeat nuclease domains.

5.2.1.2 Identification of SNP changes in clinical strains with a polymorphic SPN23F12470 gene

Table 5.1 below lists each of the SNP changes identified in the SPN23F12470 gene in the collection of clinical isolates used for GWAS. The position of each SNP in the genome is detailed, along with the SNP change, the resulting amino acid change, and the strains in which these mutations are present.

Table 5.1. SNP changes in the SPN23F12470 gene of clinical isolates. Details of each SNP change in the SPN23F12470 gene in the collection of clinical isolates used for GWAS, including the position of each SNP in the genome, the specific nucleotide change, the resulting amino acid change, and the strains containing each SNP.

Position	SNP change	Amino acid change	Strain(s) containing SNP
1,215,012	G to A	Leucine to Phenylalanine	McGeeDG104
1,216,008	C to T	Aspartate to Asparagine	McGee1789
1,216,701	C to T	Alanine to Threonine	HKP1
1,216,985	G to A	Threonine to Methionine	11923, 11919 and 11924

5.2.1.3 SPN23F12470 resides on an integrative-conjugative element

This protein resides on the integrative-conjugative element ICE*Sp*23FST81, which has clinical significance because it confers resistance to the antibiotics chloramphenicol and tetracycline. It is a composite element with homology to the transposons Tn*5252* and Tn*916* and it can be excised from the chromosome to form a circularised plasmid of around 81kb in size (Croucher et al., 2009), as depicted in figure 5.3 below.

To characterise the specific function of this UvrD-like protein in relation to toxicity in *Streptococcus pneumoniae*, a whole-gene deletion mutant was constructed as described in section 5.2.2 below. The toxicity phenotype of this mutant is characterised in detail, and the effect of this mutation on expression of other virulence-related genes in investigated. Further work in section 5.2.9 also looks at the effect of the SPN23F12470 gene deletion on mobilisation of the ICE*Sp*23FST81 element.

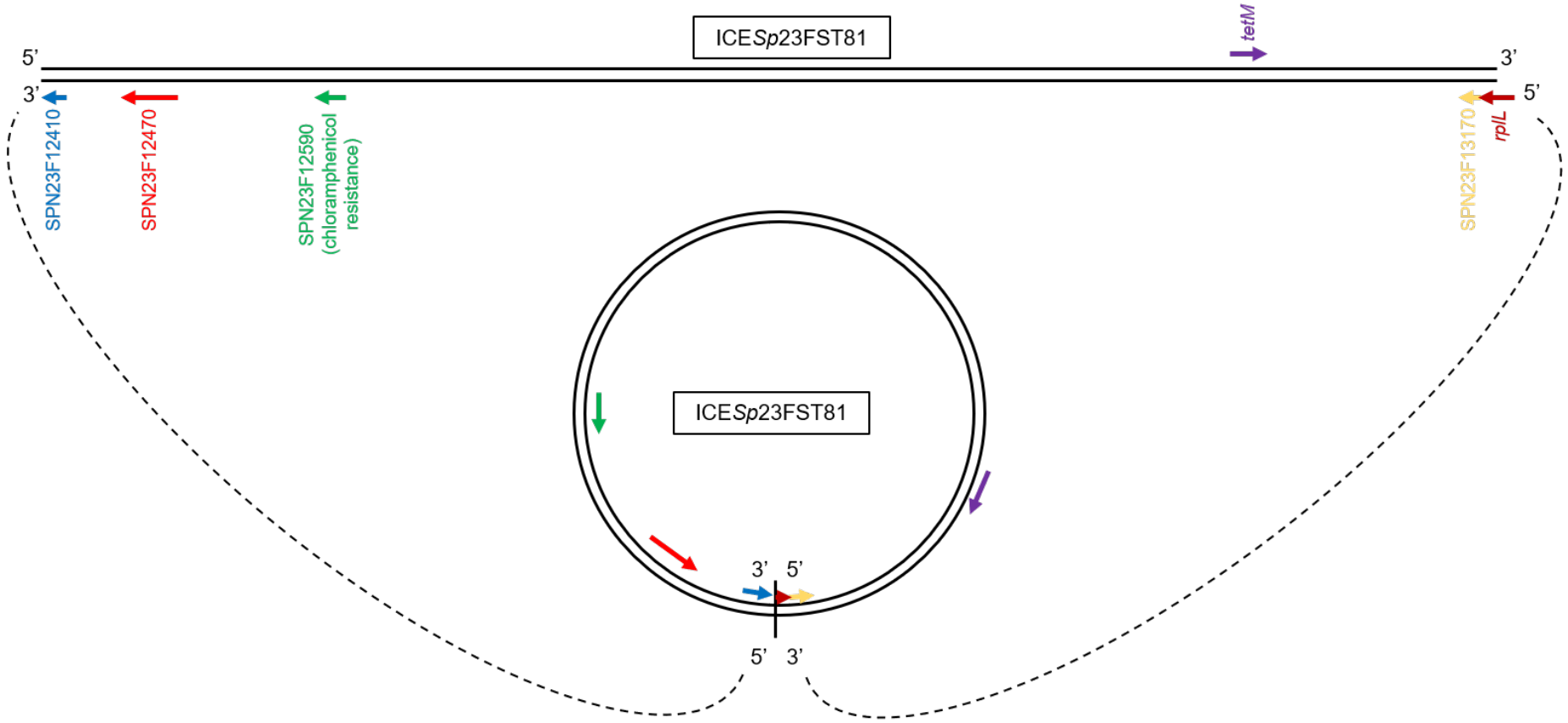


Figure 5.3. **Depiction of the integrative conjugative element ICE*Sp*23FST81**. The linear element (81,025bp in length) is excised from the chromosome and circularised into a plasmid. The element starts at gene SPN23F12410 and ends 9bp into the *rplL* gene. It contains resistance genes to the antibiotics chloramphenicol and tetracycline. The SPN23F12470 gene is shown in red.

5.2.2 Mutant construction

A ΔSPN23F12470 whole gene deletion mutant was constructed using the method described in chapter 2, section 2.3.4. Figures 5.4 & 5.5 below show the results of each PCR reaction to amplify the three fragments of the knockout insert, and the subsequent stitch PCR reaction to create the final ~2.75kb insert containing the resistance cassette and homologous regions either side.

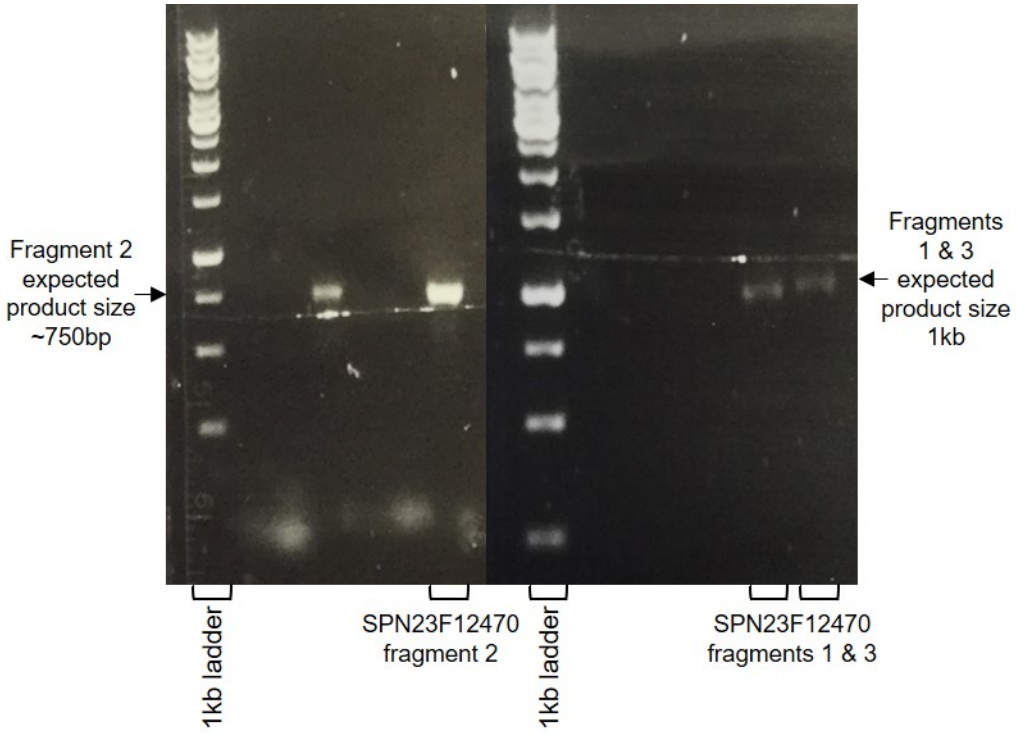


Figure 5.4. SPN23F12470 insert fragment amplification. Amplification of fragments 1-3 for construction of the SPN23F12470 insert.

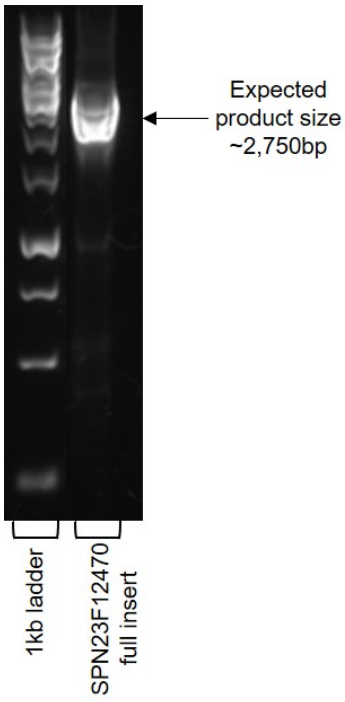


Figure 5.5. Stitch PCR for SPN23F12470 full insert. Fragments 1-3 were combined into one PCR reaction, in which the overlapping ends of each fragment were stitched together using the forward primer for fragment 1 and the reverse primer for fragment 3, creating a single linear insert around 2.75kb in length.

The ~2.75kb insert was transformed into *S. pneumoniae* strain ATCC 700669 (SPN23F) as described in chapter 2, section 2.3.4.2. Construction of the ΔSPN23F12470 mutant was verified through PCR, as shown in figures 5.6-5.8, using primers directed to:

1) The wild type *ply* gene to confirm the successful transformant was *S. pneumoniae* and not a contaminant
2) The wild type SPN23F12470 gene to verify that it was present in the isogenic wild type strain ATCC 700669 (SPN23F) and absent in the ΔSPN23F12470 mutant
3) A region starting 1kb upstream of the transformed linear insert to within the *ermAM* cassette to confirm both that the insert had inserted into the genome and that it had inserted into the correct position

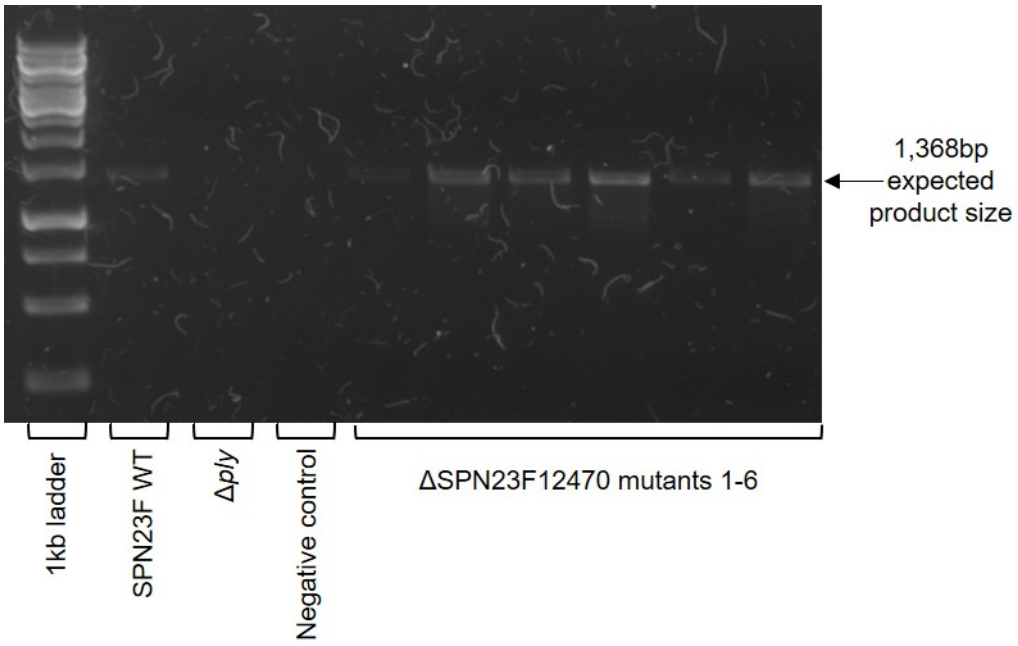


Figure 5.6. Verification that transformants are not a contaminating organism. PCR was conducted using primers to the wild type *ply* gene, to confirm that transformants were not contaminating organisms. Controls were wild type SPN23F DNA, Δ*ply* DNA and a water only negative control. ΔSPN23F12470 mutant colony 1 shows a very faint band here, which may not be clearly visible in print.

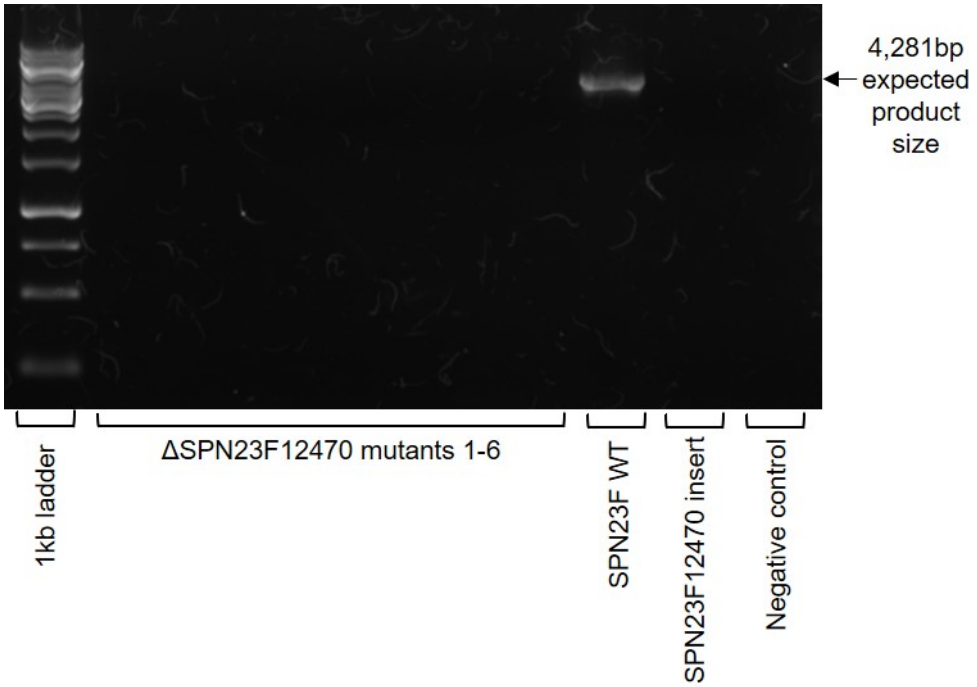


Figure 5.7. Verification that mutant colonies do not contain the wild type SPN23F12470 gene. PCR was conducted using primers to the wild type SPN23F12470 gene. Controls used were wild type SPN23F DNA, the SPN23F12470 insert used for the transformation, and a water only negative control. A band is only visible for the wild type SPN23F positive control, as expected upon successful transformation.

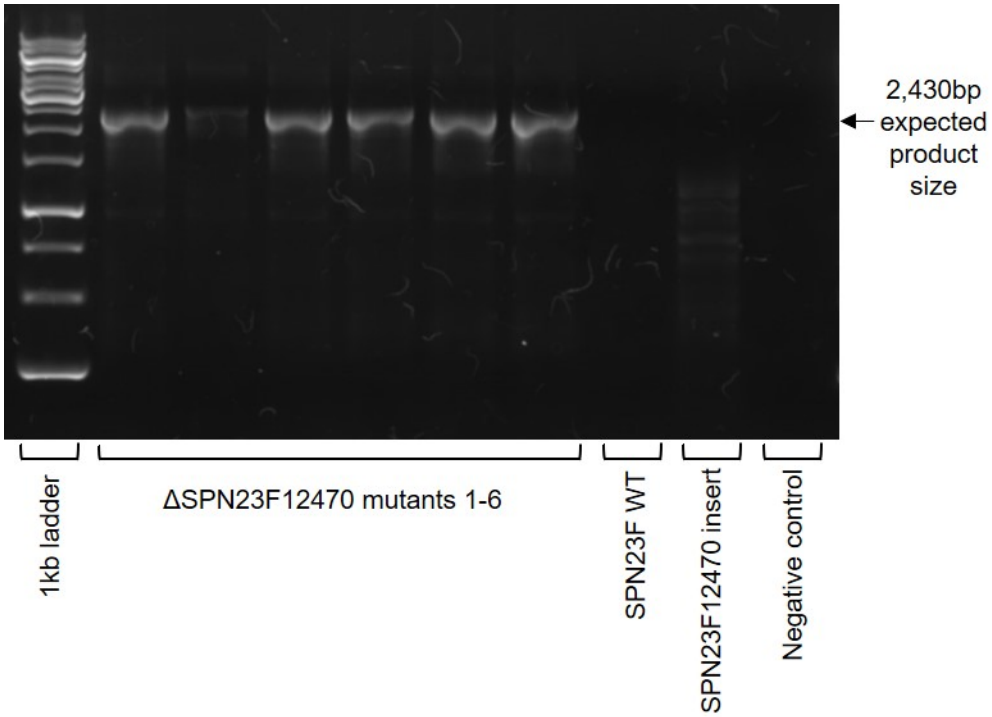


Figure 5.8. Verification that mutant colonies contain the antibiotic resistance cassette in place of the wild type SPN23F12470 gene. PCR was conducted using primers to amplify the region starting 1kb upstream of the 2.75kb transformed insert to within the middle of the *ermAM* cassette. This verifies both that the *ermAM* cassette has been taken up, and that it has recombined in the correct position within the genome to delete the wild type gene. ΔSPN23F12470 mutant colony 2 shows a faint band here at the correct size, which may not be clearly visible in print.

All subsequent assays for this mutant were conducted using the third colony because this showed clear bands for the first and third PCR checks (figures 5.6 & 5.8). Successful mutant construction was further verified for this colony through sequencing of the PCR product from figure 5.8, which confirmed insertion of the linear construct into the genome at the correct location. The primers used for sequencing were as follows:

F1: TGCCCACTATTTTTATCTAGTTGCTTACC
F2: TATCTGATCTAAGACACTTAGCAGAGG
WT F3: GGATAGATTAGAAAAATTAATTGGAG
WT RV: AAATACTTTGATCATCATCCCCTACAAC
ΔSPN23F12470 F3: GGATTCTACAAGCGTACCTTGGATATTC
ΔSPN23F12470 RV: CACCATTACGAATATTATAGAG

5.2.3 The ΔSPN23F12470 mutant shows a significant reduction in toxicity compared to the isogenic wild type

Pneumolysin is an intracellular toxin which lacks an N-terminal secretory signal sequence and is specifically localised to the wall of the pneumococcal cell. For release of the toxin into the extracellular environment, it is thought that pneumolysin is mostly dependent on the action of the autolytic enzyme LytA, which stimulates lysis of a subset of cells in a growing pneumococcal population to release intracellular pneumolysin.

The results of a toxicity assay for this mutant using whole cell culture have been presented in chapter 4, figure 4.16, but are included in figure 5.9 below for comparison. Due to the unusual nature of release of pneumolysin, a further two toxicity assays were conducted to investigate its toxic phenotype in detail. The first of these was conducted in same manner as the assay used in chapter 4, section 4.2.3, but used only the supernatant from a stationary phase culture. This was to determine specifically the level of pneumolysin which had been released into the extracellular environment. Results from this and the original toxicity assay are presented below in figure 5.9 and demonstrate a significant reduction in toxicity of ΔSPN23F12470 in comparison to the isogenic wild type using both whole cell culture and supernatant only.

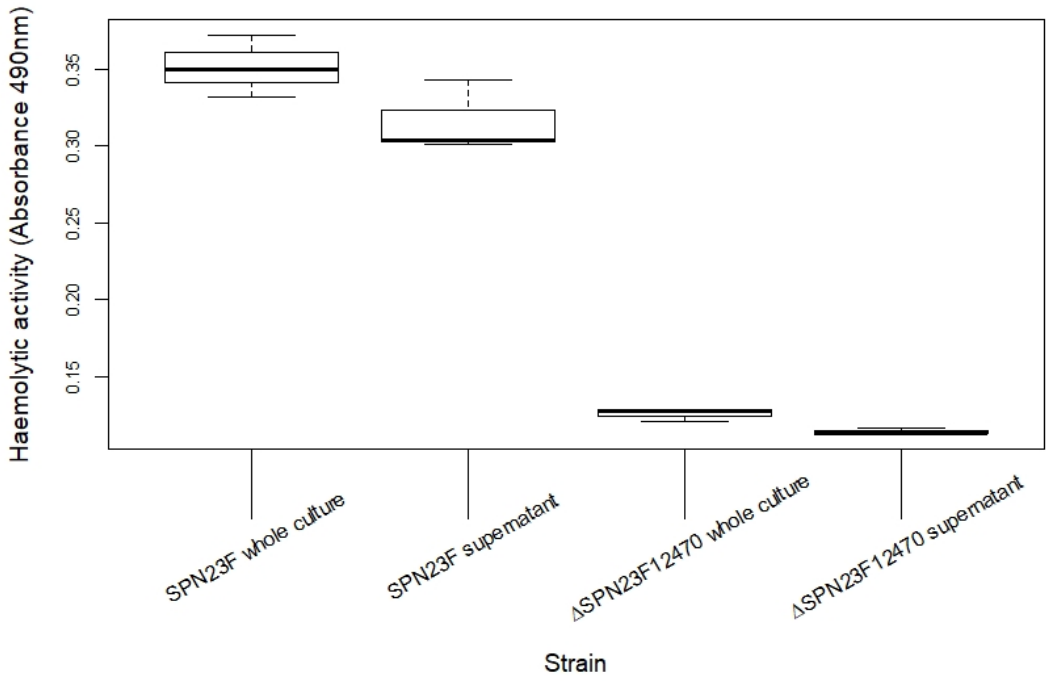


Figure 5.9. Haemolytic activity of ΔSPN23F12470 mutant whole cell culture and supernatant. A significant reduction in toxicity is observed between whole cell culture of the ΔSPN23F12470 mutant compared to the isogenic wild type strain (p=0.001) and between the supernatant only of the mutant compared to the wild type (p=0.004).

The third toxicity assay looked to investigate the effect of the ΔSPN23F12470 mutation on localisation of pneumolysin to the cell wall. This assay involved harvesting cells from a stationary phase culture for centrifugation with 25% sheep red blood cells and is referred to as a contact-dependent haemolytic activity assay. Through centrifugation of pneumococcal cells with red blood cells, the two are forced into contact with each other before being incubated for a longer period of three hours. Samples were then centrifuged again, and absorbance readings taken for the resulting supernatant, as with the previous assays. As demonstrated by higher absorbance readings for the wild type strain ATCC 700669 (SPN23F) in figure 5.10 below, toxicity of the wild type strain appears much greater when the bacterial cells are forced into contact with red blood cells and incubated for a longer period with a higher concentration of red blood cells. Because of these differences in experimental procedure, results from this experiment are presented separately to those from the previous two toxicity assays.

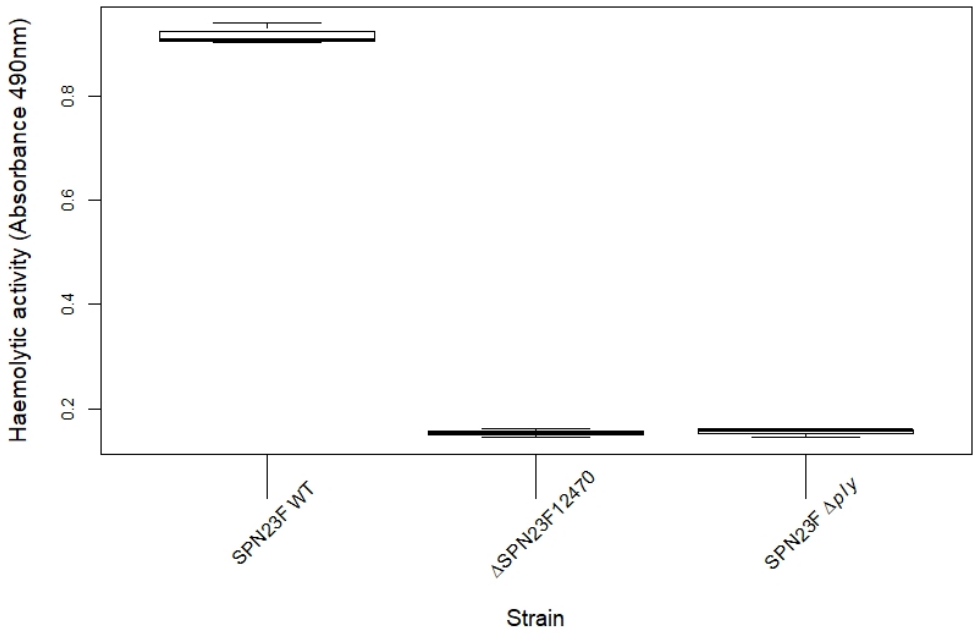


Figure 5.10. Contact-dependent haemolysis of ΔSPN23F12470 mutant. A significant reduction in contact-dependent toxicity is observed in the ΔSPN23F12470 mutant in comparison to the isogenic wild type strain ($p<0.0001$), which is comparable to the loss of toxicity observed in the Δ*ply* mutant.

This assay can be used as an initial investigation into the localisation of pneumolysin to the cell wall, with reduced toxicity here indicating less pneumolysin is in contact with red blood cells, and by proxy that there is less present in the cell wall of the ΔSPN23F12470 mutant.

5.2.3.2 Construction of a ΔSPN23F12470 mutant containing a spectinomycin resistance cassette for complementation

To complement the ΔSPN23F12470 mutant, the plasmid pMSP7517 was provided to us by Angela Nobbs at the University of Bristol. This plasmid was chosen on the basis that it contains a nisin-inducible promoter which provides control over expression of the complementation gene. As pMSP7517 is selected for using erythromycin, the original mutant was re-constructed using the spectinomycin resistance cassette *aad9*. It was later established that plasmid pVA838 could have been used for complementation, however, this plasmid does not have an inducible promoter region for control of gene expression. Toxicity data is shown in figure 5.11 for this re-constructed mutant to demonstrate a comparable loss in toxicity to the original ΔSPN23F12470 mutant using whole cell culture.

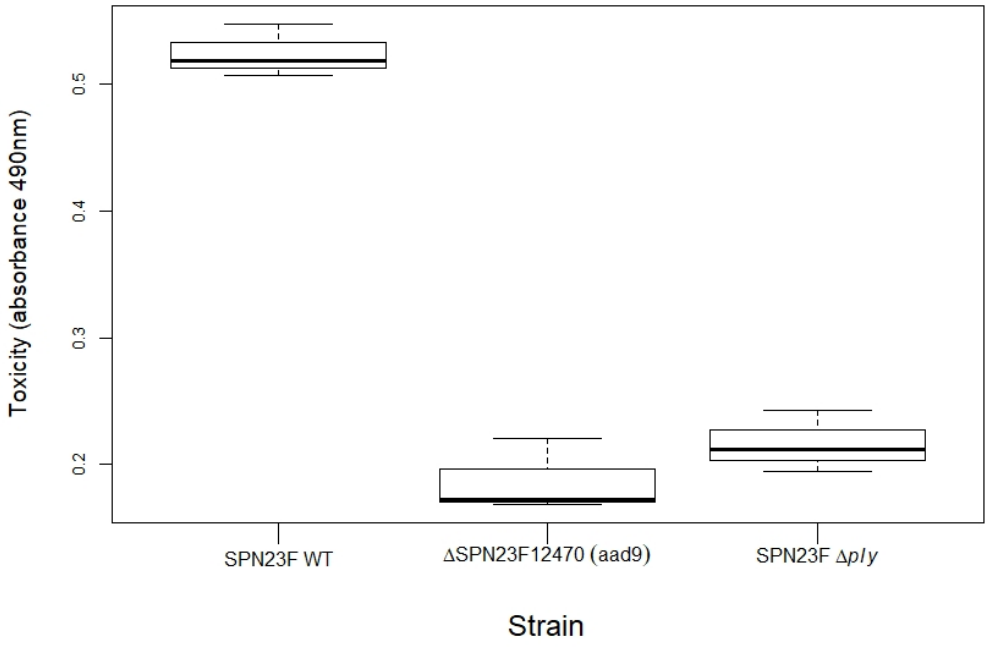


Figure 5.11. Toxicity of ΔSPN23F12470(*aad9*) whole cell culture. The haemolytic activity assay was repeated with whole cell culture for the reconstructed ΔSPN23F12470 mutant containing the spectinomycin resistance cassette *aad9* instead of *ermAM*. Toxicity is significantly reduced compared to the isogenic wild type (p=0.00017).

Unfortunately, attempts to ligate the wild type SPN23F12470 gene into pMSP7517 and transform it into the ΔSPN23F12470(*aad9*) mutant have thus far been unsuccessful, so it has not yet been possible to see if the toxicity phenotype of this mutant can be rescued with the wild type SPN23F12470 gene.

5.2.4 Loss of the SPN23F12470 gene affects pneumolysin production

To characterise the effect of the ΔSPN23F12470 mutation on pneumolysin production and release, a Western blot was conducted using protein extracted from both cell pellet and supernatant samples. Pneumolysin does not have an N-terminal secretory signal sequence, and as such is not actively secreted out of the cell. Release of this toxin is thought to be partly reliant on the autolytic activity of *S. pneumoniae* from mid-exponential to stationary phase growth, in which a proportion of the population will autolyse. This is an interesting and unusual phenomenon which, although not fully understood, has been heavily linked to virulence in the literature (Berry et al., 1989; Canvin et al., 1995; Martner et al., 2009, Skovbjerg et al., 2017). For this reason, proteins were extracted from both

the cell pellet and supernatant of ΔSPN23F12470 and its isogenic wild type, to determine if pneumolysin production inside the cell and/or its release into the extracellular environment were affected by this mutation.

Results showed a 21.56-fold reduction in the amount of pneumolysin present in the ΔSPN23F12470 cell pellet compared to the wild type, and a 26.63-fold reduction in pneumolysin in the supernatant of ΔSPN23F12470 compared to the wild type, as quantified using ImageJ and averaged from three biological repeats. Figure 5.12 below illustrates this reduction in pneumolysin quantity; here pneumolysin appears to be completely absent from the ΔSPN23F12470 mutant, comparable to the Δ*ply* mutant which was included as a control.

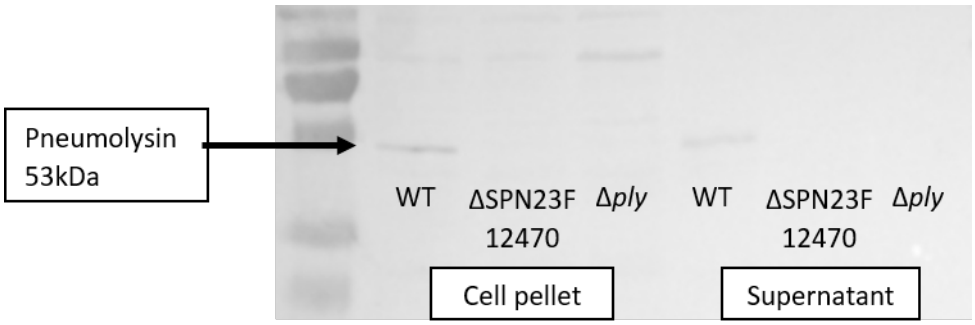


Figure 5.12. Western blots using anti-pneumolysin antibody. A substantial reduction is observed in the amount of pneumolysin (53kDa size) present in the cell pellet and supernatant protein extractions of the ATCC 700669 (SPN23F) wild type (WT), ΔSPN23F12470 and Δ*ply*, which was included for comparison.

These results imply a significant loss in production of pneumolysin in ΔSPN23F12470 which appears comparable to that of the Δ*ply* mutant. It is noted however that cell pellets of mutants which are deficient in autolytic activity are resistant to the detergent lysis solution used here to obtain cell pellet extractions (Balachandran et al., 2001). As such, although a significant difference is observed here in the amount of pneumolysin between wild type and mutant, it is possible that a deficiency in autolysis could be affecting this result; if pneumolysin is present in the ΔSPN23F12470 cell pellet and the mutant is autolysis-deficient, cells will not have lysed and pneumolysin will not be evident in the cell pellet extraction from ΔSPN23F12470 in this blot. Deoxycholate could have been used here to lyse bacterial pellets, however, to specifically show whether or not autolysis is affected in the mutant, an autolysis assay was subsequently conducted.

5.2.5 The ΔSPN23F12470 mutant is not deficient in autolytic activity

Autolysis of a proportion of the pneumococcal population is thought to be mostly responsible for the release of pneumolysin into the extracellular environment, since the pneumolysin protein does not contain an N-terminal secretory signal sequence for exportation out of the cell. For detailed description of the mechanism of autolysis and its proposed association with toxicity, see chapter 1, section 1.4.2.3. Autolytic activity of ΔSPN23F12470 was therefore tested (figure 5.13) to demonstrate both that deficiency in autolysis is not affecting the results of the Western blot above, and that it is not the reason for the observed loss of toxicity in the haemolytic activity assays.

For this assay, a ΔSPN23F15290 mutant was used as a control, as this gene is annotated as an autolysin in strain ATCC 700669 (SPN23F) and a mutant in this gene had been constructed following its significant association with toxicity from the SEER results (chapter 4, section 4.2.2.2). This gene does not encode the main autolysin LytA, which is encoded by the *lytA* gene, however, deletion of SPN23F15290 yields an autolysis-deficient phenotype and therefore this mutant was a suitable control to use.

The effects of autolysis can be measured *in vitro* by growing the bacteria to stationary phase, harvesting and washing the cells and resuspending them in PBS to equivalent optical densities. Triton X-100 is added to a final concentration of 0.1%, after which cells are incubated at 37°C and $OD_{(600)}$ readings taken at 15 minute intervals to quantify detergent enhanced autolysis. Figure 5.13 below shows the results expressed as percentage cells remaining alive, using the average of three biological repeats for each strain and using data from time point 0 to represent 100%. For this reason, there are no error bars for the first time point.

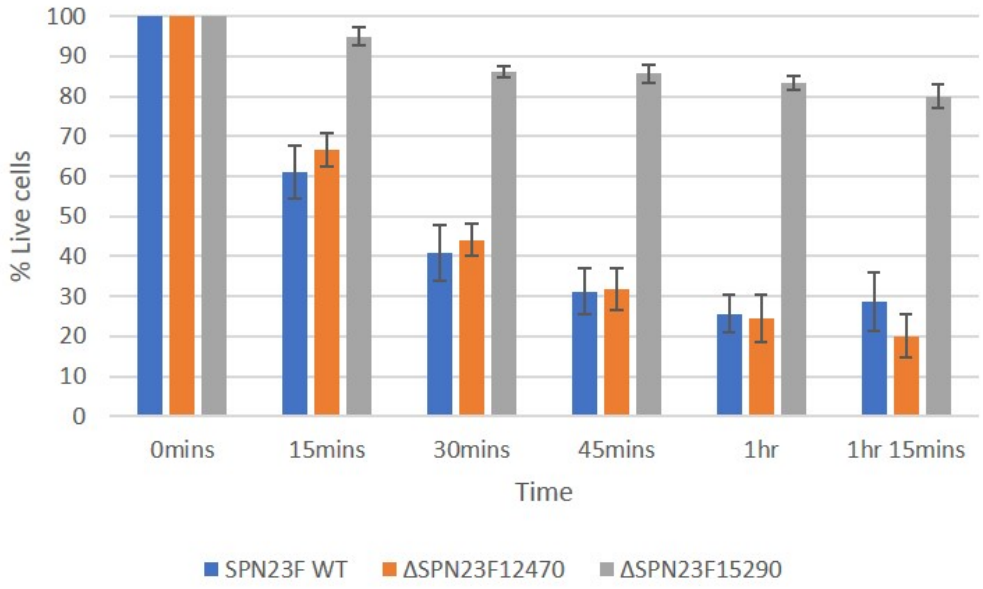


Figure 5.13. The ΔSPN23F12470 mutant is not deficient in autolytic activity. Autolytic activity was measured by assaying the lysis of a stationary phase culture following addition of 0.1% Triton X-100. Data are expressed as percentage cells remaining intact and the difference in % live cells between the wild type and the ΔSPN23F12470 mutant is not significant.

Using this assay, a clear difference was observed in cell density between the autolysin-deficient ΔSPN23F15290 mutant compared to the wild type and ΔSPN23F12470 mutant strains, whilst there was no significant difference between the wild type and ΔSPN23F12470. This confirms that autolysis is not affected in the ΔSPN23F12470 mutant, as cells in this culture lysed comparably to the isogenic wild type strain, and critically it demonstrates that the loss of toxicity in ΔSPN23F12470 is not due to a deficiency in autolytic activity.

5.2.6 ΔSPN23F12470 exhibits reduced transcription of the *ply* gene

As the results of the toxicity assays and Western blotting have demonstrated a reduction in pneumolysin production in ΔSPN23F12470, and as this has been shown not to be due to a deficiency in autolytic activity, qRT-PCR was used to investigate whether this effect on toxicity was at the transcriptional level. qRT-PCR was conducted to the *ply* and *lytA* genes, with the major autolysin gene *lytA* included here for verification that autolysis is not affected in the ΔSPN23F12470 mutant. The competence genes *comA* and *comE* (annotated as gene SPN23F22680) were also included to investigate whether there is a change in

the competence phenotype of the mutant in comparison to the wild type, as competence is thought to be linked to virulence in the pneumococcus. It was hypothesised that with such a strong impact on toxicity in the ΔSPN23F12470 mutant, competence could also be affected.

The *comA* gene encodes a subunit of the ComAB transporter complex which exports the competence-stimulating peptide (CSP) out of the pneumococcal cell and *comE* is a DNA-binding protein that regulates expression of genes in the competence operon. For further description of the competence operon in *S. pneumoniae* and its relation to pneumococcal virulence, see chapter 1, section 1.4.2.5.

RNA was extracted from two cultures each for the ΔSPN23F12470 mutant and the isogenic wild type strain and converted to cDNA through reverse transcription. cDNA quantity was standardised prior to addition to the qPCR reaction mix, and three technical repeats were conducted for each cDNA sample. For further details of the method, see chapter 2, section 2.3.5.6. Results are shown in figure 5.14 below as fold-change in transcript expression of genes in the ΔSPN23F12470 mutant compared to the isogenic wild type strain, as calculated by comparison to the housekeeping gene *recA*, using the $2^{-\Delta\Delta CT}$ method. Error bars represent standard error of the mean.

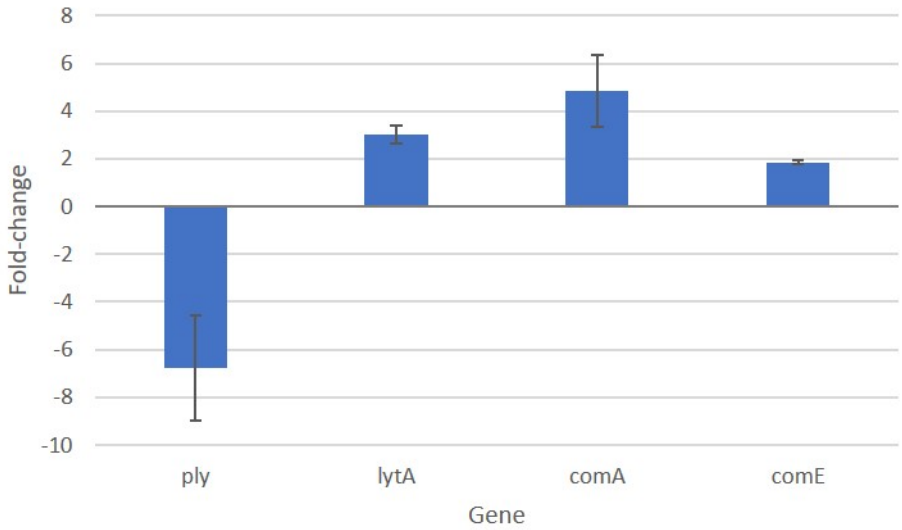


Figure 5.14. qRT-PCR results. Fold-change in transcript expression of key virulence and competence-related genes in the ΔSPN23F12470 mutant compared to the isogenic wild type, as calculated by comparison to the housekeeping gene *recA*, using the $2^{-\Delta\Delta CT}$ method. Results are based on two biological repeats.

These results support previous findings from the toxicity assay and Western blots, showing that transcript expression for the *ply* gene is down-regulated 6.8-fold in the ΔSPN23F12470 mutant. It also supports the finding of the autolysis assay, showing that there is no reduction in autolytic activity in the ΔSPN23F12470 mutant; in fact a 3-fold increase is observed here in transcript expression for *lytA*, which could explain the slightly lower percentage of cells remaining alive in the ΔSPN23F12470 mutant compared to the wild type at the 1hr 15mins time point in figure 5.13 above. A 4.8-fold increase is also observed for the *comA* gene, suggesting that transcript expression of competence machinery is up-regulated in ΔSPN23F12470, however, transcript expression of *comE* is not up-regulated to the same extent, with only a 1.8-fold increase observed here which implies very little change to transcription of this gene.

5.2.7 The UvrD-like protein encoded by SPN23F12470 binds to regions either side of the *ply* gene containing BOX repeats

Having identified that loss of the SPN23F12470 gene affects transcription of the *ply*, *lytA* and *comA* genes, it was observed that BOX-type repeat regions were present in the neighbouring intergenic regions beside each of these genes. BOX repeat regions are a family of short repetitive fragments of DNA between 45-59 base pairs in length that are present at over 100 different positions within the pneumococcal genome and have been shown to affect transcription of neighbouring genes (Knutsen et al., 2006). They usually consist of a combination of boxA (59bp), B (45bp) and C (50bp) repeat sequences, which can overlap, be in different orientations and in different combinations throughout the genome. Generally, boxA and boxC flank the boxB region, which can be repeated more than once and be in different orientations to the flanking repeat regions. Regions containing boxA and boxC elements are also known to form stable stem-loop structures, which could be involved in modulating the expression of neighbouring genes (Knutsen et al., 2006).

In the case of the *ply* gene, BOX repeats are found immediately downstream in the neighbouring intergenic region. They are also found 5-7 genes upstream, however these upstream repeat regions are only one gene distant from the start of the *ply* operon, as annotated in strain D39V in the online programme PneumoBrowse (Slager et al., 2018). The downstream BOX repeat regions which

immediately neighbour *ply* are found in the order AB_1CB_2, with all repeat sequences in the same orientation as the two neighbouring genes, encoded on the lagging strand (see figure 5.15).

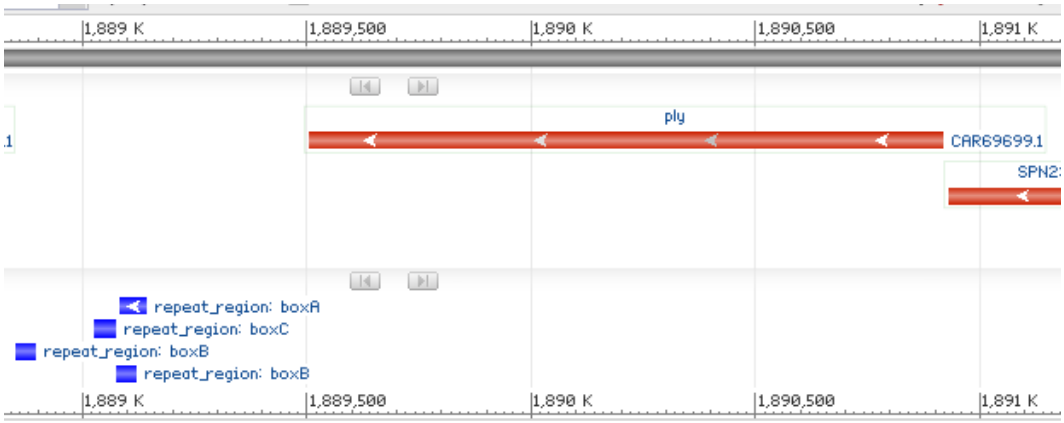


Figure 5.15. BOX repeat regions downstream of the *ply* gene in strain ATCC 700669 (SPN23F). A screenshot of the *ply* gene from the FM211187.1 annotation of the genome on the National Centre for Biotechnology Information website and its downstream intergenic region in which four BOX repeat regions can be found in the combination AB_1CB_2. All are encoded on the lagging strand like the two neighbouring genes.

Based on the presence of a helicase region and two nuclease domains within the UvrD-like protein encoded by SPN23F12470, and based on the down-regulation of *ply* in the mutant as shown by qRT-PCR, it was hypothesised that the UvrD-like protein could be affecting transcription of the *ply* gene. Furthermore, it is known that Cas proteins, to which the nuclease domains of this UvrD-like protein have homology, bind to repeat sequences which guide the protein to a target region of DNA; therefore it was thought that interactions between the SPN23F12470 protein and the BOX repeat regions could be mediating the effect on transcription of *ply*. To investigate this, an electrophoretic mobility shift assay (EMSA) was conducted in which specific binding of a protein to small regions of DNA can be visualised. For this assay, wild type UvrD-like protein was purified by Oliver Wilkinson in the Dillingham lab (see chapter 2, section 2.3.5.8 for method details).

Firstly, the intergenic region downstream of *ply* containing the BOX repeats was amplified through PCR, yielding a product size of 882bp, and a region of corresponding size was amplified from the *ply* gene itself for use as a control. The downstream repeat region was chosen initially because of the closer proximity of these repeat regions to the *ply* gene. An EMSA was then conducted

using a range of protein concentrations and a fixed DNA concentration of 10nM to determine whether the wild type protein would bind to either section of DNA. Further details of this method can be found in chapter 2 section 2.3.5.8.

Results shown in figure 5.16 indicate binding of the UvrD-like protein to the intergenic region containing BOX repeats downstream of *ply* at the lower protein concentrations of 31.25nM and 62.5nM, as demonstrated by a shift in size of the DNA band at these concentrations, to which protein is bound. High-affinity binding is visualised by a distinct, clear band at a higher position on the gel than the no protein control, while smeary bands indicate non-specific binding, as is observed at higher protein concentrations and for the *ply* region of DNA.

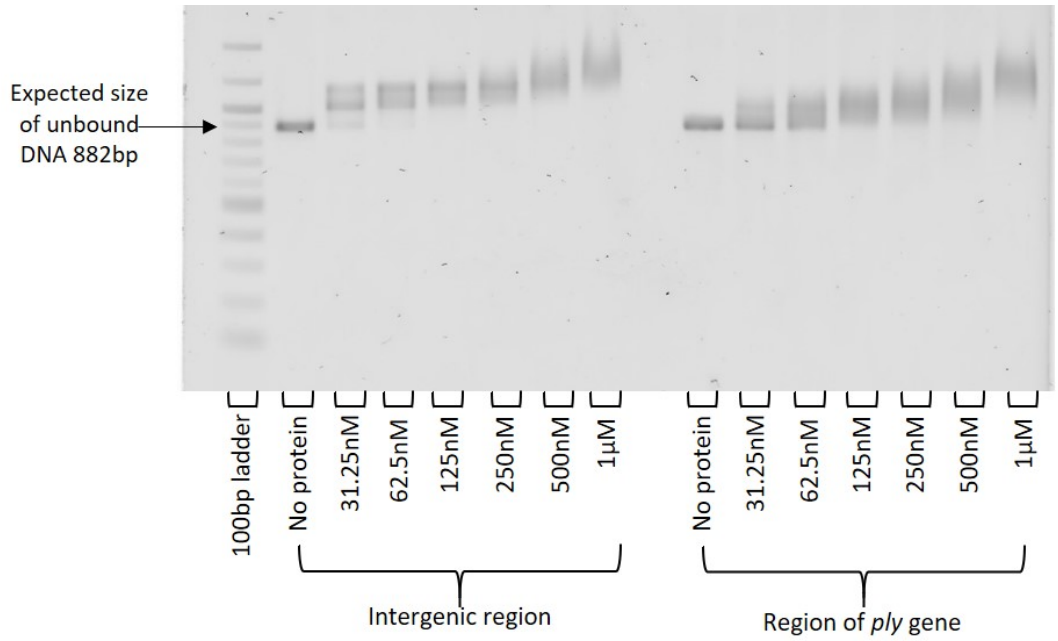


Figure 5.16. EMSA showing high-affinity binding of the UvrD-like protein to the intergenic region containing BOX repeats. A range of protein concentrations from 0-1µM were used initially to determine at what concentration range, if any, the protein would bind to DNA. Results show high-affinity binding of the protein to the intergenic region downstream of *ply* containing BOX repeats, at a concentration range of 31.25-62.5nM protein, as demonstrated by the presence of a distinct band of bigger size than the no protein control at these concentrations.

To further clarify this result, the EMSA was repeated with a smaller, more focused range of protein concentrations and a lower concentration of DNA (5nM compared to 10nM in the previous experiment). Figure 5.17 shows a clear shift of the intergenic region containing BOX repeats at all concentrations, but most

distinctly at concentrations of 2.5-10nM protein, while the *ply* control shows non-specific shifting from 20-80nM protein.

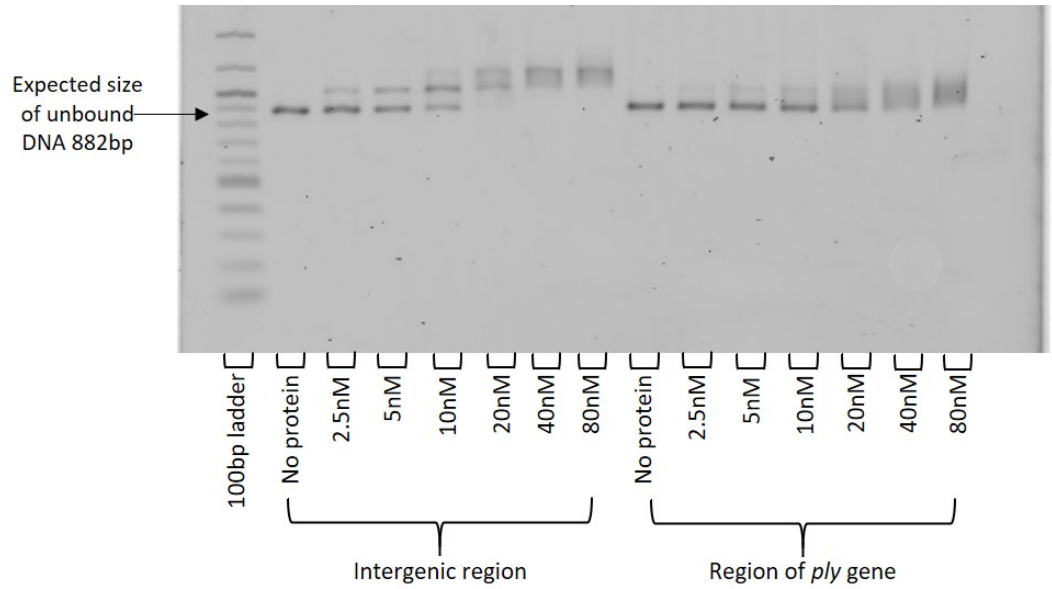


Figure 5.17. EMSA repeat using lower range of protein concentrations. Results demonstrate clear, specific binding of the UvrD-like protein to the region of DNA containing BOX repeats, while there is mostly no binding or some non-specific binding of the protein to the *ply* DNA.

Results of these EMSAs show high-affinity binding of the UvrD-like protein to the intergenic region downstream of *ply* containing BOX repeats. However, the concentration of DNA used in both experiments (10nM and 5nM respectively) was higher than the lowest concentration of protein used (2.5nM). This means the protein is saturated at this point, and as such substrate binding at this concentration cannot be accurately determined. To address this, DNA was radiolabelled with phosphorous-32 and the EMSA repeated, such that a DNA concentration of 2nM could be used to obtain an accurate picture of substrate binding at the lowest protein concentration of 2.5nM. For this assay, an 882bp region of DNA was included from the region 2,758bp upstream of *ply* which also contains BOX repeats. The up- and down-stream BOX repeat regions used as substrates here correspond to the regions either side of the *ply* operon, as annotated in strain D39V in the online programme PneumoBrowse (Slager et al., 2018). Figure 5.18 shows the result of the first of two repeats of this assay using 2nM radiolabelled DNA.

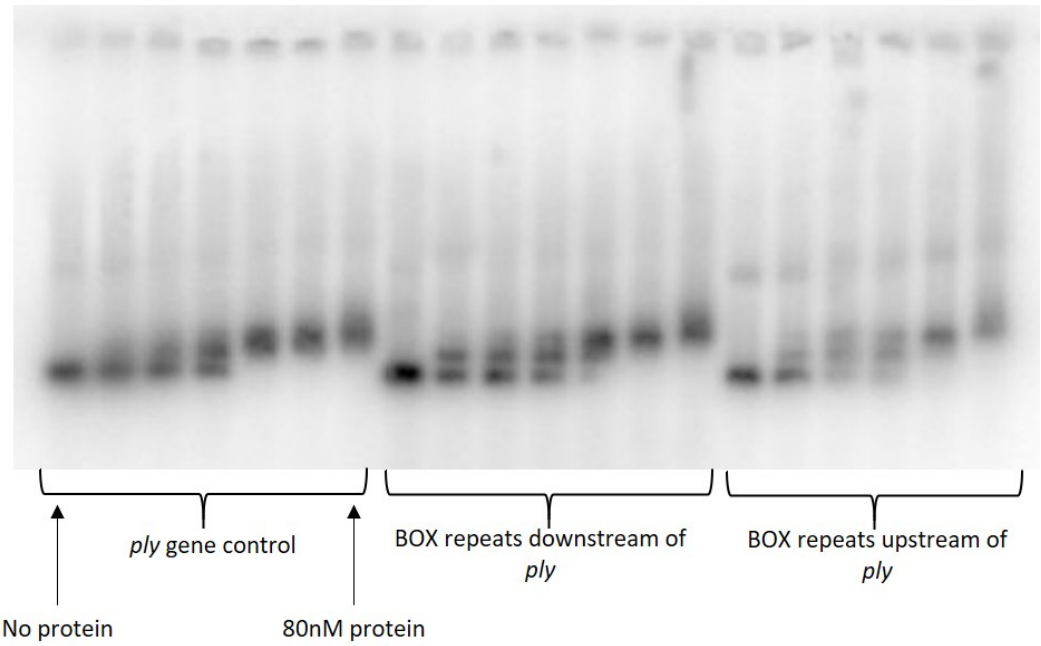


Figure 5.18. EMSA repeat using 2nM DNA labelled with phosphorous-32. 882bp regions of DNA from the *ply* gene, the BOX repeat region downstream of *ply* and an additional BOX repeat region upstream of *ply* were radiolabelled and added to a final concentration of 2nM in this EMSA repeat. Protein concentrations were 0, 2.5, 5, 10, 20, 40 and 80nM.

Following a second repeat of this assay, disappearance of the lowest band (corresponding to the size of the no protein control) as protein concentration increases was quantified using ImageJ and converted to represent percentage substrate to which the protein is bound. The bands which were quantified are indicated by the red rectangle in the inset image of figure 5.19. This data is then presented in figure 5.19 as a substrate binding curve from 0-20nM protein, at which point 99-100% of each substrate is bound to the protein.

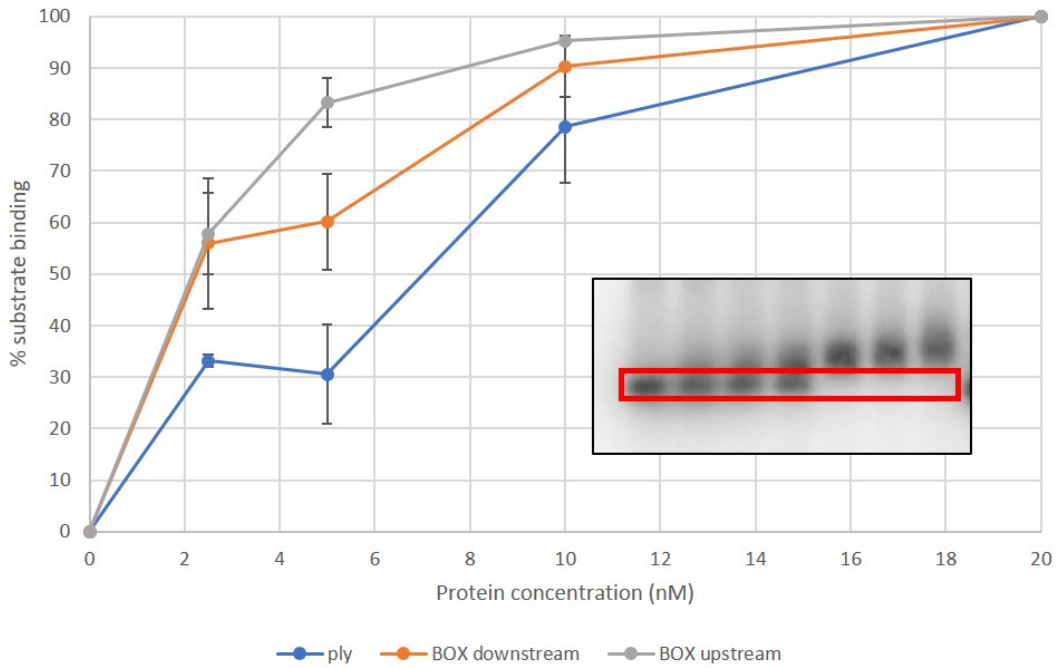


Figure 5.19. Substrate binding curves. Percentage substrate binding for each region of DNA was quantified by measuring the reduction in intensity of the bands corresponding in size to the no protein control, as indicated in the inset figure here. This measurement was then converted to represent the percentage of substrate bound to the protein. Data is shown for protein concentrations from 0-20nM, as 99-100% of each substrate was bound to the protein at a concentration of 20nM.

These results demonstrate a physical interaction between the UvrD-like protein encoded by SPN23F12470 and regions of the genome either side of the *ply* operon that contain BOX repeats, suggesting that it may be through interactions with these repeat regions that the UvrD-like protein exerts its effect on transcription of the *ply* gene.

5.2.9 The ICE*Sp*23FST81 element is present in higher quantities in circularised form in the ΔSPN23F12470 mutant

After demonstrating a physical interaction between the UvrD-like protein encoded by SPN23F12470 and regions of the genome neighbouring the *ply* gene containing BOX repeats, it was observed (using the FM211187.1 annotation of the genome on the National Centre for Biotechnology Information website) that BOX repeat regions are present two genes upstream of the *rplL* gene at the end of the ICE*Sp*23FST81 element. It was therefore hypothesised that the UvrD-like protein could be involved in mobilisation of the ICE*Sp*23FST81 element via an interaction with these BOX repeat regions. To test this hypothesis, PCR was

conducted to the junction of the circularised form of the element where the two linear ends join, using a forward primer for gene SPN23F12410 and a reverse primer for gene SPN23F13160 as shown in figure 5.20 below (see chapter 2, section 2.3.5.5 for method details).

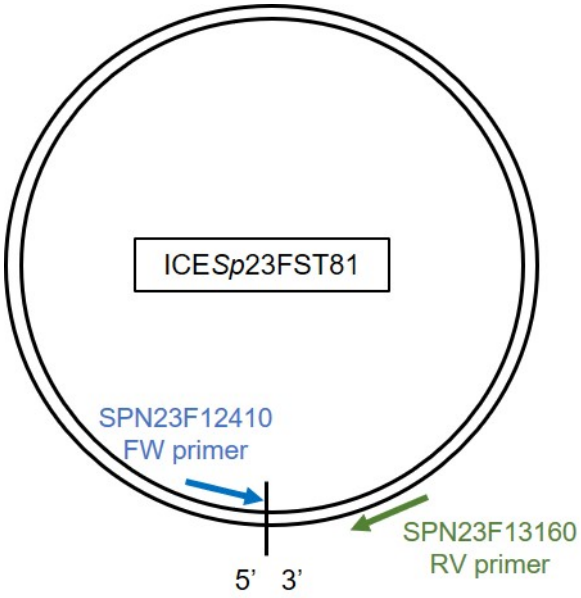


Figure 5.20. Primers used to amplify circularised ICE*Sp*23FST81. A forward primer for gene SPN23F12410 was used along with a reverse primer for gene SPN23F13160 to amplify the junction where the two linear ends of ICE*Sp*23FST81 join.

Sequencing of this PCR product for the junction of the circularised ICE*Sp*23FST81 element (extracted from the ATCC 700669 SPN23F wild type) where the two linear ends meet revealed that it starts within an intergenic region at the 3' end of gene SPN23F12410, which is 6 genes distant from SPN23F12470, and ends 9 base pairs into the 3' end of the *rplL* gene, which is 70 genes distant from SPN23F12470. Figure 5.21 below shows the sequence where the two linear ends join.

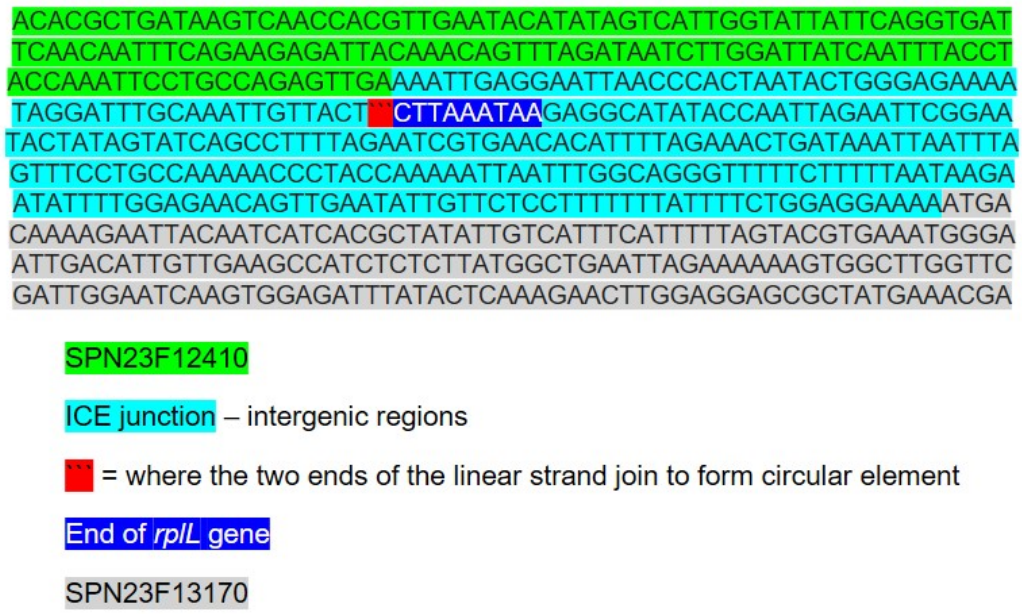


Figure 5.21. Sequencing results for the region where the linear ends of ICE*Sp*23FST81 join. Sequencing of the junction of the circularised ICE*Sp*23FST81 element shows the two linear ends join at an intergenic region 57 bases downstream of the 3' end of the SPN23F12410 gene and 9 bases into the 3' end of the *rplL* gene.

After establishing that the circularised form of the element was present in both the wild type ATCC 700669 (SPN23F) strain and the ΔSPN23F12470 mutant, PCR was conducted on three separate plasmid extractions from each strain to determine whether there was a difference in quantity of the circularised element. DNA concentration was standardised such that the final concentration in each PCR reaction was 20ng. Figure 5.22 shows consistently higher quantities of the circularised form of this ICE*Sp*23FST81 element in the ΔSPN23F12470 mutant compared to the isogenic wild type.

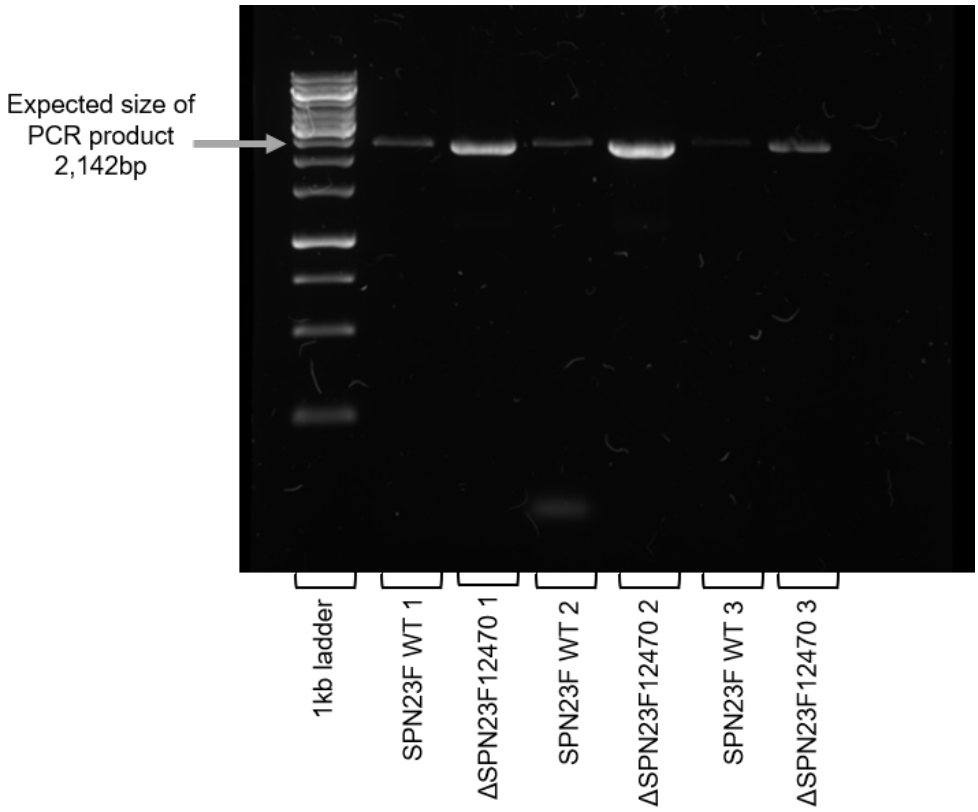


Figure 5.22. PCR to detect the circularised form of ICE*Sp*23FST81. Circularised ICE*Sp*23FST81 is present in higher quantities in the ΔSPN23F12470 mutant, as demonstrated by PCR to the junction where the two linear ends of the element join, using plasmid extractions from three biological repeats for each strain.

These results suggest that the circularised form of the ICE*Sp*23FST81 element is present in higher quantities in the ΔSPN23F12470 mutant compared to the isogenic wild type; this could be either due to increased excision of the ICE element from the chromosome in the mutant or decreased uptake of the circularised element into the chromosome of the mutant.

5.3 Discussion

Results presented here demonstrate a clear link between activity of the UvrD-like protein encoded by the SPN23F12470 gene and production of the pneumolysin toxin of *S. pneumoniae* in strain ATCC 700669 (SPN23F). Toxicity assays show a significant reduction in *in vitro* toxicity of the ΔSPN23F12470 mutant compared to its isogenic wild type and Western blotting confirms reduced production of the pneumolysin protein. Autolytic activity of the mutant was assayed in comparison to the isogenic wild type and an autolysis-deficient strain, confirming that loss of toxicity in the mutant was not related to a deficiency in autolytic activity, which is corroborated by qRT-PCR data to the major autolysin gene *lytA*. qRT-PCR data

further demonstrates that the loss in production of pneumolysin is due to a reduction in transcript expression of the *ply* gene in ΔSPN23F12470. Up-regulation of the *comA* gene is also shown through qRT-PCR, implying an increase in expression of competence machinery in ΔSPN23F12470, although this has not been confirmed phenotypically.

Analysis of the UvrD-like protein encoded by SPN23F12470 revealed that the gene encodes a DNA helicase domain homologous to the *uvrD* gene found in *E. coli*, with two tandem repeat active nuclease domains at the N-terminus with homology to the Cas4 nuclease of the CRISPR-Cas system. From communication with collaborator Mark Dillingham at the University of Bristol, it was suggested that the SNP change observed in clinical strain McGeeDG104 from leucine to phenylalanine (table 5.1) is within a highly conserved locus, positioned just outside the amino acid region which would normally be defined as helicase motif VI, which is important for ATP hydrolysis and coupling hydrolysis to DNA translocation. The change from aspartate to asparagine in strain McGee1789 is suggested to occur within a somewhat conserved region close to helicase motif IV, but asparagine can be found at this locus in some cases anyway. The change from alanine to threonine in strain HKP1 takes place in an insert region that is only present in extremely similar proteins, so the level of conservation of this sequence cannot be established. The change from threonine to methionine in strains 11923, 11919 and 11924 is within a poorly conserved locus, but is close to the Q motif, otherwise known as motif 0, which is important for ATP binding and nucleotide hydrolysis specificity. All these insights have come from personal communication with Mark Dillingham, who also suggested that it is likely some of these mutations lead to problems in ATPase function.

However, as shown in figure 5.1, clinical isolates with a polymorphism in the SPN23F12470 gene have on average higher toxicity than those with the reference SPN23F12470 gene. This could mean that these polymorphisms are instead hyper-activating and lead to an improvement in the function of the UvrD-like protein; toxicity data indicates that in the absence of this protein toxin production is significantly reduced, suggesting that it normally serves to promote toxin production in the wild type strain.

Based on the observation of the presence of BOX repeat regions neighbouring the *ply*, *lytA* and *comA* genes, for which transcript expression is affected, a potential interaction between the UvrD-like protein and regions of the genome

containing BOX repeats was investigated, with a specific focus on the *ply* gene to elucidate whether it via this interaction that the UvrD-like protein is exerting its effect on toxicity. It was demonstrated that the protein interacts with high affinity with the regions either side of *ply* containing BOX repeat regions. BOX repeat regions have been shown in work by Knutsen et al. (2006) to influence expression of their neighbouring genes. They are considered parasitic mobile genetic elements that appear to have inserted randomly and at regular intervals throughout pneumococcal genomes in which they are present (Croucher et al., 2009). It is also known that BOX repeat regions containing both boxA and boxC elements can form secondary stem-loop structures (Knutsen et al., 2006).

In the wild type strain ATCC 700669 (SPN23F), BOX repeat regions downstream of *ply* are present in the intergenic region between the stop codon of *ply* and the start codon of its downstream neighbour. They are encoded on the lagging strand in the same orientation as the two flanking genes and they are in the combination AB_1CB_2. BOX repeats are also found 5-7 genes upstream of the *ply* gene prior to the start of the *ply* operon (as annotated in strain D39V in the programme PneumoBrowse [Slager et al., 2018]).

Based on the high-affinity binding of the UvrD-like protein to regions of DNA containing these BOX repeats, while a control region of the *ply* gene does not show the same degree of affinity, it is possible that in the wild type ATCC 700669 (SPN23F) strain, the UvrD-like protein serves to unwind and/or degrade the secondary structures formed by these repeats through the action of its helicase and nuclease domains. This interaction with BOX repeat regions up- and downstream of *ply* could be important for normal expression of the wild type *ply* gene, which could explain how this protein affects toxin production in this *S. pneumoniae* strain.

To test this hypothesis, further work would be needed to investigate the direct link between the UvrD-like protein, the BOX repeat regions and expression of the *ply* gene. This could be done by deleting the BOX repeat regions upstream and downstream of *ply* in the ΔSPN23F12470 mutant. As this mutant contains an intact *ply* gene but lacks the UvrD-like protein, it would be expected that deletion of the BOX repeat regions either side of *ply* would remove this physical block to transcription, allowing normal expression of pneumolysin to proceed. Single deletions of the upstream and downstream BOX repeat regions could be done

alongside a double deletion of both repeat regions to characterise whether one region alone or both regions in tandem are affecting transcription of the *ply* gene.

Interestingly, a region of BOX repeats is also found upstream of the *comA* gene, for which an increase in transcription is observed in the ΔSPN23F12470 mutant. It could be the case that in the wild type strain, SPN23F12470 serves to increase expression of the pneumolysin gene through its interaction with the neighbouring BOX repeat regions, while simultaneously inhibiting expression of competence machinery. The upregulation of competence machinery in ΔSPN23F12470 could also be related to the observed increase in mobilisation of the mobile genetic element ICE*Sp*23FST81, on which the SPN23F12470 gene is located.

ICE*Sp*23FST81 is an integrative conjugative element present in strains of the SPN23F ST81 lineage, although it has regions of homology to transposons found in other strains. From a clinical perspective, acquisition of this element provides several benefits to the host, as it contains both chloramphenicol and tetracycline resistance cassettes, and based on the findings presented here, a gene which serves to promote pneumolysin production in its host. The absence of any BOX repeat regions within the ICE*Sp*23FST81 element, in spite of their otherwise regular appearance on average every 17.3kb in this strain, implies this element integrated into the genome of ATCC 700669 (SPN23F) after the BOX repeat regions. Therefore, it is particularly interesting that the ICE element encodes a gene which interacts with regions of the genome containing BOX repeats to the apparent benefit of the host.

The presence of higher quantities of circularised ICE*Sp*23FST81 element within the ΔSPN23F12470 mutant in comparison to the isogenic wild type also implies that loss of this gene causes either greater excision of the ICE element from the chromosome or reduced ability to reintegrate the element back into the chromosome. This could be a direct result of the disruption of the SPN23F12470 gene, or it could be an indirect consequence whereby loss of SPN23F12470 affects other genes involved in excision or reintegration of the element into the chromosome or its transmission out of the cell.

Interestingly, recent work from several groups has identified a link between CRISPR-Cas systems and bacterial virulence (Louwen et al., 2014; Faure et al., 2019; Ratner et al., 2019). Clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR-associated (Cas) genes have been identified as a bacterial system of adaptive immunity, whereby small CRISPR RNA sequences

guide Cas nucleases towards a complementary nucleic acid target sequence of invasive genetic elements such as viruses and plasmids. Cas nucleases then work to degrade the invading DNA to prevent infection of the chromosome by potentially harmful genetic elements. Mobile genetic elements have been shown to contribute to the origin and evolution of CRISPR-Cas systems, which are thought to have evolved from a specific type of transposon described as a 'casposon', which encodes a transposase enzyme with homology to the Cas1 nuclease (Faure et al., 2019). Transposase enzymes catalyse the integration of mobile genetic elements into the genome, and one of the best characterised classes of transposase enzymes is the DD(D/E) transposases, so named based on the amino acid residue of their catalytic domain (Faure et al., 2019; Guérillot et al., 2014). Examination of the amino acid sequence for the UvrD-like protein revealed a highly conserved DDD residue at amino acid positions 388-390, which could imply a potential role for this protein as a transposase. Indeed, there is a possibility that the nuclease activity of this protein is causing the difficulty in complementing the whole gene deletion mutant, as it may be fatal to the *E. coli* host that is used to passage the vector containing the wild type gene. Attempts were made to clone the wild type gene directly into the pneumococcus, but these were unsuccessful.

Faure et al. (2019) suggest that the casposon transposase has been co-opted for use in bacterial adaptive immunity, and further highlight that some casposons encode a Cas4 nuclease, to which the two tandem repeat nuclease domains of the UvrD-like protein have homology. It is therefore tempting to speculate that the protein encoded by SPN23F12470 is involved in transposition of the ICE*Sp*23FST81 element, and furthermore that it is involved in mediating the effects of the parasitic BOX repeat elements in strain ATCC 700669 (SPN23F) through degradation of the secondary structures that they form via the Cas4-like nuclease domains. This could be a form of adaptive immunity against these invasive genetic elements that have integrated throughout the genome.

5.3.1 Further work

Time constraints prevented this work being taken further, however, detailed mechanistic characterisation of the link between the UvrD-like protein and pneumolysin production could be done by looking in more detail at the connection between the protein, the BOX repeat regions and the *ply* gene. For example, it would be interesting to investigate whether the protein is binding BOX

repeat regions all over the genome and if this is how it affects, for example, competence gene expression. Interactions between the protein and other areas of the genome could be investigated using chromatin immunoprecipitation-sequencing assays (ChIP-seq), in which, briefly, cells containing a tagged UvrD-like protein would be grown to a desired stage, then the protein and any regions of DNA to which is it bound would be extracted and the bound DNA purified and sequenced. This would elucidate specific regions of DNA to which the wild type UvrD-like protein binds. Further biochemical analysis of the function of the UvrD-like protein would also help to elucidate whether this protein is in fact acting as a transposase comparable to the bacterial CRISPR-Cas adaptive immunity system.

Complementation of the ΔSPN23F12470 mutant with the wild type SPN23F12470 gene has thus far been unsuccessful. Attempts to complement the mutant will be continued, to show that toxicity in ΔSPN23F12470 can be restored with re-introduction of the wild type gene. Work is ongoing to do this in the ΔSPN23F12470 mutant, and in parallel, attempts will be made to introduce the wild type SPN23F12470 into strain D39, to see if an increase in toxicity is observed in a strain that does not contain this protein, but does contain BOX repeats downstream of the *ply* gene.

It would also be of interest to further investigate the relationship between the UvrD-like protein and competence, with a focus on how the protein affects mobilisation of ICE*Sp*23FST81. It would be particularly interesting to investigate whether there is a direct link between competence and toxicity in this strain which is mediated by the UvrD-like protein, since much is known about the relationship between competence and autolysis, and likewise autolysis and toxicity. There could be a direct link between competence and toxicity, which have thus far been indirectly linked via the mechanism of autolysis.

5.4 Conclusions

To conclude, loss of the SPN23F12470 gene causes a significant reduction in toxicity in *S. pneumoniae* strain ATCC 700669 (SPN23F) as shown using whole cell culture, supernatant and contact-dependent haemolytic activity assays. A 6.8-fold reduction is observed in transcript expression of the *ply* gene in the ΔSPN23F12470 mutant, causing a significant reduction in pneumolysin production which is confirmed by Western blotting. The ΔSPN23F12470 mutant is not deficient in autolytic activity, as shown through a phenotypic assay and

qRT-PCR, the latter of which also shows that transcript expression of the *comA* gene is increased 4.9-fold. Perhaps related to this, presence of the mobile genetic element ICE*Sp*23FST81 in circularised form is increased in the ΔSPN23F12470 mutant in comparison to its isogenic wild type. These results are summarised in figure 5.23 below.

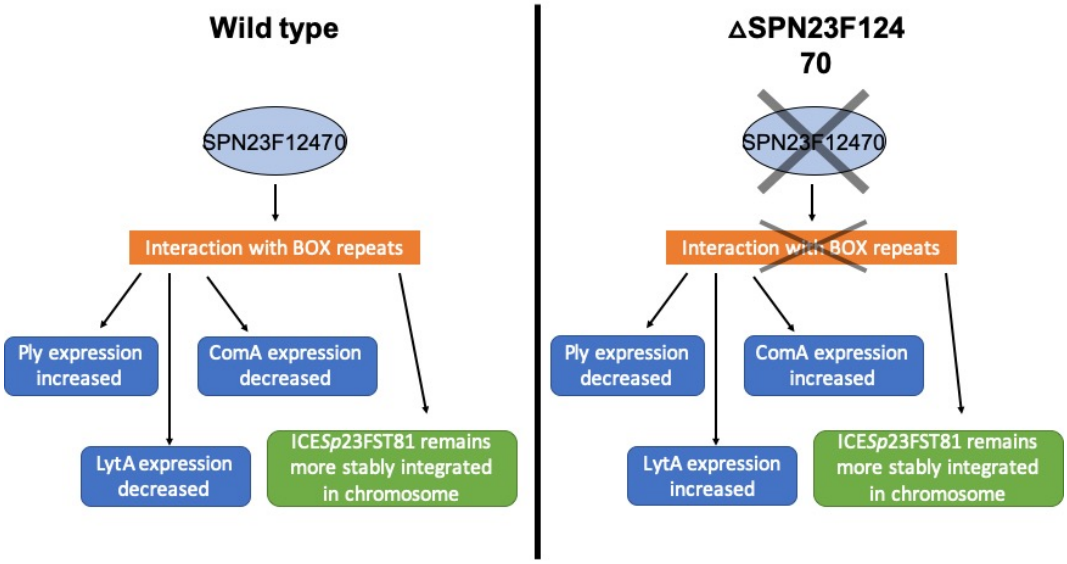


Figure 5.23. Summary of findings. In the presence of the wild type SPN23F12470 gene, via interactions with neighbouring BOX repeat regions, transcript expression of pneumolysin is upregulated while transcript expression of autolysin and competence machinery is downregulated and the ICE*Sp*23FST81 element remains more stably integrated in the genome. In the absence of the SPN23F12470 gene, toxicity is reduced due to downregulation of *ply*, while transcript expression of autolysin and competence machinery is upregulated, and the ICE*Sp*23FST81 element is found in its circularised form at comparably higher quantities relative to the wild type.

It is shown that the UvrD-like protein encoded by the SPN23F12470 gene binds with high affinity to the intergenic region immediately downstream of *ply* which contains BOX repeats and the region upstream of *ply* also containing BOX repeats. Research by Knutsen et al. (2006) has illustrated that BOX repeat regions can affect the transcription of neighbouring genes, and that they form stable secondary stem-loop structures. A possible explanation therefore for the relationship between the UvrD-like protein encoded by SPN23F12470 and toxicity in this strain could be that under normal circumstances the protein interacts with the BOX repeat regions neighbouring *ply* to promote its transcription by unwinding and degrading the structures formed by these repeat regions. In the absence of this protein, the BOX repeats may form a physical block to transcription of *ply*, preventing successful expression of the gene and resulting in the subsequent reduction in toxicity. There is a possibility the UvrD-

like protein is acting as a form of adaptive immunity against these invasive BOX repeat elements that have integrated throughout the pneumococcal genome.

These results demonstrate the important role mobile genetic elements can play in bacterial virulence, and how spread of such elements can have a dramatic effect on hosts into which they integrate. Identification of this UvrD-like protein as a genetic contributor to toxicity is a novel finding which contributes to the body of knowledge regarding the genetic effectors of pneumolysin production in *Streptococcus pneumoniae*. It also demonstrates how the top-down GWAS approach to identifying novel effectors of a phenotype can be successful, and how this approach can help to expand our knowledge of individual gene function in relation to bacterial virulence.

6. Conclusions

The overarching aim of this work was to investigate the genetic basis for bacterial toxicity using functional genomics. Toxicity is an important aspect of virulence which is often responsible for the symptoms of bacterial disease in a patient. It is regulated by a complex network of genetic interactions which is not yet fully understood. The work presented here has focussed on identifying and characterising novel toxicity-affecting loci in two clinically important pathogens; *Staphylococcus aureus* and *Streptococcus pneumoniae*. By further elucidating genetic contributors to the toxic phenotype in these bacteria, this work adds to the body of research which looks to better understand a key aspect of bacterial virulence.

In *S. aureus*, follow-up work was conducted from a prior genome-wide association study (GWAS) to characterise the involvement of the *cyoE* gene in staphylococcal toxicity. This identified the protoheme IX farnesyltransferase enzyme encoded by *cyoE* as a critical component in toxicity, without which the cell down-regulates TCA cycle activity, respires as if growing in microaerobic conditions and reduces toxin production and secretion through down-regulation of the global toxicity regulator *agr*.

This work has highlighted the important link between bacterial metabolism and virulence, demonstrating how disruption of a gene which plays a crucial role in electron transport can have a profound downstream effect on the energetically costly process of toxin production. The fact that single nucleotide polymorphisms in the *cyoE* gene in clinical *S. aureus* strains exhibit reduced toxicity also corroborates the importance of metabolic processes to toxin production *in vivo*. As such, identification of genes such as *cyoE* and its resulting protein product protoheme IX farnesyltransferase, which is not in itself a virulence factor yet contributes to the phenotype, may provide an avenue for the development of future therapies against *S. aureus*.

Following the successful identification of a novel toxicity-associated gene in *S. aureus*, the functional genomics approach was applied to a collection of clinical *S. pneumoniae* isolates. In this work, three different genome-wide association methods were applied to the same dataset to aid prioritisation of genes for further investigation. Of a total of 181 genes and intergenic regions that were identified

by the three methods, 8 genes were identified by more than one method. Due to difficulties encountered in the mutant construction process, functional verification was not possible for each of these 8 genes, however, 2 were identified as false positive associations; *priA* and *trcF*. A further 6 genes which were each identified by only one GWAS method were also identified as false positive associations.

However, despite the small number of individual loci which were identified by more than one method as significantly associated with toxicity, neighbouring genes within clusters of the genome were picked up by combinations of the three different methods. For example, the SNP-based GWAS method and Bugwas identified genes in the region from *agaD* to *clpL* as significant associations, while the SNP-based method and SEER overlapped in identification of genes on the ICE*Sp*23FST81 element, and all three methods identified genes around the *priA* locus.

Problems encountered with the process of functional verification, due in part to the background strain chosen for mutant construction, have limited the extent to which loci can be verified as false positive or true positive associations with the phenotype. Therefore, firm conclusions as to which method, or combination of methods, is most accurate in predicting genetic associations with toxicity cannot be drawn at this stage. Differences in the initial data filtration steps, controlling of population structure, and significance thresholds are likely to account for the differences in results obtained by the three methods, however it is particularly interesting that no single gene has been identified as a significant association with toxicity by all three methods.

Nonetheless, one gene was confirmed to be a true association with pneumococcal toxicity; this was the SPN23F12470 gene, encoding a UvrD-like protein with one helicase and two tandem-repeat nuclease domains. Further investigation of the contribution of this protein to toxicity has illustrated its importance in transcription of the pneumolysin gene and has suggested a role for the protein in competence and mobilisation of ICE*Sp*23FST81 in strain ATCC 700669 (SPN23F). Initial data implies a physical interaction between the UvrD-like protein and the BOX repeat regions neighbouring the pneumolysin gene which may be affecting expression of the toxin. In the absence of the UvrD-like protein, these BOX repeat regions may form a physical block to transcription of the operon on which *ply* is encoded, which would otherwise be unwound and degraded through the action of SPN23F12470.

Investigation of the differences in quantity of circularised ICE*Sp*23FST81 between the wild type strain and the ΔSPN23F12470 mutant demonstrate a higher quantity of the circularised mobile genetic element in ΔSPN23F12470, suggesting either increased excision of the element in the mutant or reduced re-integration into the chromosome. This observation, in combination with both the observed interaction of the UvrD-like protein and the BOX repeat regions neighbouring *ply* and the homology of its two tandem repeat nuclease domains with Cas4 proteins, could imply a role for SPN23F12470 similar to the CRISPR-Cas system of bacterial adaptive immunity. The UvrD-like protein could be using BOX repeat regions to guide its activity of unwinding and degrading target sections of DNA. It could also be important in this capacity for transposition of the ICE*Sp*23FST81 element either out of or into the host chromosome, in which case this protein may be important in mediating transmission of ICE*Sp*23FST81 between hosts.

Given the observed positive effect of the UvrD-like protein on toxin expression, and therefore virulence, in the wild type strain ATCC 700669 (SPN23F), it is interesting that the ICE*Sp*23FST81 element on which this protein is located has not propagated throughout other pneumococcal lineages; perhaps there is a fitness cost to incorporation of this element into the host genome. It is also interesting that SPN23F12470 is shown here to interact with regions of the genome containing BOX repeats, which integrated into the chromosome prior to the integration of ICE*Sp*23FST81. Furthermore, whilst interactions with specific regions containing BOX repeats neighbouring *ply* are investigated here, it would be of interest to investigate whether the UvrD-like protein also interacts with BOX repeat regions throughout the genome of this strain to mediate other aspects of bacterial behaviour.

Both *S. aureus* and *S. pneumoniae* are commensal organisms with the ability to transition to opportunistic pathogens within the human host. Whilst the genes encoding specific toxins produced by these two organisms have been characterised, the work presented here illustrates how complex the genetic regulation of toxicity can be, with identification of a metabolic gene in *S. aureus* and a gene encoded on a mobile-genetic element in *S. pneumoniae* which have both been observed to significantly impact toxin production when disrupted. Despite the stark differences in the number of different toxins employed by these bacteria, this work has emphasised that regardless of species, toxin production in

bacteria is affected by a wide-ranging network of genetic contributors for which our understanding is very much incomplete.

The functional genomics approach used to investigate the genetic basis for toxin production, whilst it has limitations, has proven successful in identification of a novel effector of toxicity in both species studied here. Without the statistical associations made in the GWAS analyses used, both the *cyoE* and SPN23F12470 loci may have gone unnoticed as significant contributors to the toxic phenotype. Furthermore, this approach is useful in the first instance to establish which genes are involved in bacterial toxicity, but it is also useful to understand how SNP changes in these genes in clinical isolates can impact toxin production *in vivo*. Understanding the impact these loci have on toxicity in isolates causing disease may allow for better interpretation of genetic information should it ever come to be used in diagnosis of bacterial infections. For example, the ability to predict the likely severity of disease caused by an infecting isolate will allow for tailored treatment to the needs of the individual patient.

The use of three different GWAS methods on the same dataset showed that, unexpectedly, there was very little crossover of loci that were identified as statistically significantly associated with toxicity between the three methods. Due to the small number of results that could be functionally verified it is not possible at this stage to say whether one method is more accurate in its prediction of toxicity-associated loci than the other two. Nonetheless, this approach led to the identification of a second novel association with toxicity; the SPN23F12470 gene. Whilst this gene is annotated as a putative DNA helicase II with homology to UvrD, the results presented here suggest that this is not a complete picture of the protein's function. Work is ongoing to elucidate this function, with a specific focus on the role of the protein in pneumococcal toxicity, which like *cyoE*, had not been previously identified.

In the long term, it is hoped that research such as this will collectively lead to improved prevention and treatment of disease caused by *S. aureus* and *S. pneumoniae* through a better understanding of pathogenesis in these bacteria. With the substantial advances made in genome sequencing technology over recent years, and the increased availability of genome sequence data, work such as this demonstrates how a genomic approach can be used to study one of the fundamental phenotypes which contributes to bacterial virulence.

Appendix I

I.I Primers used to construct whole gene deletion mutants in the background strain ATCC 700669 (SPN23F):

Primer	Sequence	Tm
ply F1	CCCTTGCTCTGGTTAAAAAAAGAAGC	61.6
ply R1	ATATTTTTGTTCATATTTGCCATCTTCTACC	60.2
ply F2	GGTAGAAGATGGCAAATATGAACAAAAATATAAAA	61.3
ply R2	CTACCTGAGGTTATTTCCTCCCGTT	63
ply F3	GAGGAAATAACCTCAGGTAGAAGATAAG	62.2
ply R3	GATCACCTTTTTTAGCTGCTACATAG	60.1
nanA F1	CTAATTTAGAGAATGAAATATAGAG	53.1
nanA R1	ATTTTTGTTCATCAGATTTCTCCTTTATATTG	59.2
nanA F2	GAAATCTGATGAACAAAAATATAAAATATTCTC	58.3
nanA R2	TATTTATTATTATTTCCTCCCGTTAAATAATAG	58.3
nanA F3	GAGGAAATAATAATAAATAAAATCAATATG	54.5
nanA R3	ATGCAATAAAGTACATAAAGATTTG	53.1
priA F1	CTGGTCCTTCAGGGGTTGGAAAAGG	66.3
priA R1	ATTTTTGTTCATTAGGGCCATCGTCTCAC	63.9
priA F2	GATGGCCCTAATGAACAAAAATATAAAATATTC	60.8
priA R2	CATGGTCAATGCTTTATTTCCTCCCGTTA	63.9
priA F3	GAGGAAATAAAGCATTGACCATGAGCCGC	66.7
priA R3	CACAAAAACATCCTCTAGCACTGCTAGAG	65.3
SPN23F12470 F1	TGCCCACTATTTTTATCTAGTTGCTTACC	62.4
SPN23F12470 R1	ATTTTTGTTCATTGTTGTCATCGTTTTACCTC	61.8
SPN23F12470 F2	CGATGACAACAATGAACAAAAATATAA	55.9
SPN23F12470 R2	CTTTTCCGGATTCTTATTTCCTCCC	61.3
SPN23F12470 F3	GGAGGAAATAAGAATCCGGAAAAG	59.3
SPN23F12470 R3	AATTAATTCCTGAATACAAGTTAACAAAATAG	58
SPN23F01610 F1	ATCACAGCGGTTAAGGATGTGAC	60.6
SPN23F01610 R1	ATTTTTGTTCATTGATTCCATCTTACTGTAC	60.2
SPN23F01610 F2	GATGGAATCAATGAACAAAAATATAAAATATTCTC	60.1
SPN23F01610 R2	GTCAAGAAATCTTATTTCCTCCCGT	59.7
SPN23F01610 F3	GGGAGGAAATAAGATTTCTTGACTAAG	60.4
SPN23F01610 R3	ACCATAGGTCAAAACTAGACCCAC	61
SPN23F01620 F1	TAAAGTATTTAT AGA TTC CCT CAA TAC GAG	59.9
SPN23F01620 R1	ATATTTTTGTTCATTTTCTTCATGAATTGTCTCC	61
SPN23F01620 F2	CAATTCATGAAGAAAATGAACAAAAATATAAAATATT C	59.7
SPN23F01620 R2	TTACTCAAGGCTTATTTCCTCCCG	61
SPN23F01620 F3	GGGAGGAAATAAGCCTTGAGTAAAA	59.7
SPN23F01620 R3	CGAGAAGCAAGAAATCGTGGAAC	60.6
SPN23F02870 F1	GCGTTGCTCCGGTATGACG	61

SPN23F02870 R1	TTTTTGTTCATAATTTTCAATTTGTTTTCCTC	59.5
SPN23F02870 F2	AAACAAATTGAAAATTATGAACAAAAATATAAAATATTCTC	59.4
SPN23F02870 R2	CATTTCCAACAATTATTTCCTCCCG	59.7
SPN23F02870 F3	GGAGGAAATAATTGTTGGAAATGATATTTAC	60.2
SPN23F02870 R3	GTCATATAATGGTTTAATAAAAATCCTGATAAAAC	60.1
SPN23F15290 F1	GATGAGTTCAATTGTATCTATCGGC	59.7
SPN23F15290 R1	ATTTTATATTTTTGTTCATAATTTCCATATTCTACTCCTTATC	62.8
SPN23F15290 F2	GGAGTAGAATATGGAAATTATGAACAAAAATATAAAATATTCTC	63.8
SPN23F15290 R2	CCATCTGGCTCTACTTATTTCCTCCCGTTAAATAATAG	68.4
SPN23F15290 F3	AACGGGAGGAAATAAGTAGAGCCAGATGGC	68.1
SPN23F15290 R3	TCAAGGTATCCATCATTCCTCAATCTATATAAC	63.3
SPN23F12470 (*aad9*) F1 **	TGCCCACTATTTTTATCTAGTTGCTTACC	62.4
SPN23F12470 (*aad9*) R1	CAAATATATCCTCCTGTTGTCATCGTTTTACC	64.4
SPN23F12470 (*aad9*) F2	AAACGATGACAACAGGAGGATATATTTGAATAC	63.3
SPN23F12470 (*aad9*) R2	CTTTTCCGGATTCTTATAATTTTTTTAATCTG	59.2
SPN23F12470 (*aad9*) F3	GATTAAAAAAATTATAAGAATCCGGAAAAGAAAG	59.8
SPN23F12470 (*aad9*) R3	AATTAATTCCTGAATACAAGTTAACAAAATAG	58
trcF F1	CTGAATTGGATGATGAAGATAAAAAAGAGTT	60.2
trcF R1	ATTTTTGTTCATGGTCACCATTTATCCG	60.7
trcF F2	TAAATGGTGACCATGAACAAAAATATAAAATATTC	60.1
trcF R2	CTTAAACTCTTTTATTTATTTCCTCCCGTTAA	60.5
trcF F3	GGAGGAAATAAATAAAAGAGTTTAAGGAAG	59.9
trcF R3	AATCTTCATTCGCATAAACAGGAACC	60.1
gidB F1	AGAACCTGTCGCAAAGGCATTTAAGGC	65
gidB R1	ATTTTTGTTCATTGGTTTCATAGTTATTTCCTTCTTTAG	63.1
gidB F2	ACTATGAAACCAATGAACAAAAATATAAAATATTCTCAAAAC	62.6
gidB R2	CATACCAGCCTTTTATTTCCTCCCGTTAAATAATAG	66
gidB F3	GGGAGGAAATAAAAGGCTGGTATGCCAAATAAAC	67.1
gidB R3	ACCAGTTGGCTTTAAGAACTGGTGTAAAGTC	65.5

**The ΔSPN23F12470 mutant was reconstructed with the spectinomycin resistance cassette *aad9* for the purposes of complementation. For further details, see Chapter 5, section 5.2.3.2.

I.II Primers used to verify successful mutant construction:

Primer name	Sequence	Tm
ply FW	GCAGTAAATGACTTTATACTAGCTATGATAAAC	62
ply RV	GAGAGTTGTTCCCCAAATAGAAATCG	61.6
ply mutant check FW	ATCTCTTCAAAAGTGGTACGCTGAACACTAAC	65.6
ply mutant check RV	GTGTTCGGTGAATATCCAAGGTACGCTTG	66.7
nanA FW	ATGAAAGATTTAACTAAATACAAAGGCG	57.8
nanA RV	TTATAAAAATTCTTTTCGTGCTTGTTG	55.9
nanA mutant check FW	TCTTGCGCAATGAAGCTCATCTTTCCCTTTG	66.8
nanA mutant check RV	GTGTTCGGTGAATATCCAAGGTACGCTTG	66.7
priA FW	GCTAAGATTATCGTAGATGTGCCCTTGATG	65.4
priA RV	GAGTCAAGGCCAAGACCTGGTTGAGGGTC	70.9
priA mutant check FW	GTCGATGTGATTATCGACGATACACCAG	65.1
priA mutant check RV	GTGTTCGGTGAATATCCAAGGTACGCTTG	66.7
SPN23F12470 FW	GATTTTAAAACTTTATTTAAAGAAGAATATG	54.9
SPN23F12470 RV	GACAATAATCACAATCGTTTTTACCACAAC	61.3
SPN23F12470 mutant check FW	ATAAAAGAGCAACAAACTTTACAAATCTACTGATTG AATC	64.3
SPN23F12470 mutant check RV	GTGTTCGGTGAATATCCAAGGTACGCTTG	66.7
SPN23F01610 FW	GATTCAAACCTATTTACCAAATGTCTAT	57.8
SPN23F01610 RV	CCAAGAATTGGATTGCAAAGATAATC	58.5
SPN23F01610 mutant check FW	GATTTACCTGTCAATATCAGCTTTATCATGGAG	64.5
SPN23F01610 mutant check RV	GTGTTCGGTGAATATCCAAGGTACGCTTG	66.7
SPN23F01620 FW	ATCTATCTTAACTACACTGCTTTTTGCAG	61
SPN23F01620 RV	GTCCAATCAAGGCACCCACTAG	62.1
SPN23F01620 mutant check FW	ATGGACTTGGGATAGCGAACCATGCAG	66.5
SPN23F01620 mutant check RV	GTGTTCGGTGAATATCCAAGGTACGCTTG	66.7
SPN23F02870 FW	GGGAGAGTTTTATCGAATCTCTG	58.9
SPN23F02870 RV	GATATTCATTAAGTCATGAGAGTTTTTC	57.8
SPN23F02870 mutant check FW	ATCTCTGCAGGTGAACTTCGTTCTAAAAATG	64.2
SPN23F02870 mutant check RV	GTGTTCGGTGAATATCCAAGGTACGCTTG	66.7
SPN23F15290 FW	TCGATAGAAACAGACTACGTACAGGCTTG	65.3
SPN23F15290 RV	CTGCCATGCTACCGTCTGGTTTGAG	66.3
SPN23F15290 mutant check FW	CATGGTAGAGCTTCTGGTATTTGGACAAC	65.3
SPN23F15290 mutant check RV	GTGTTCGGTGAATATCCAAGGTACGCTTG	66.7
SPN23F12470 (*aad9*) FW	GATTTTAAAACTTTATTTAAAGAAGAATATG	54.9
SPN23F12470 (*aad9*) RV	GACAATAATCACAATCGTTTTTACCACAAC	61.3

SPN23F12470 (*aad9*) mutant check FW	ATAAAAGAGCAACAAACTTTACAAATCTAC	58.6
SPN23F12470 (*aad9*) mutant check RV	GAATATCAGGTAGTAATTCCTCTAAGTC	60.7
trcF FW	ATGGTGACCTTATTAGATTTATTCTCAG	59.3
trcF RV	TCAAATGGAATTTTCTTCCTTAAACTCTTTTATC	61
trcF mutant check FW	ATCACACGCTTGTACAATCAAGTCGACTGG	66.8
trcF mutant check RV	GTGTTCGGTGAATATCCAAGGTACGCTTG	66.7
gidB FW	CATTTTACAACTTGCTTGCCGAGCAGAATC	65.4
gidB RV	CCACTGTGATATAGCGCGGATCTCTATTC	66.7
gidB mutant check FW	AGTCTCAATTTCAGTTATCCAGAAATGCAGCATC	65.9
gidB mutant check RV	GTGTTCGGTGAATATCCAAGGTACGCTTG	66.7

Primers designated *gene_name FW/RV* are to amplify the wild type allele of the corresponding gene.

Primers designated *gene_name mutant check FW/RV* are to amplify the region from 1kb upstream of the transformed 3kb insert to within the antibiotic resistance cassette.

ply FW and RV primers to the wild type *ply* gene were used for all mutants, except Δ*ply*, to verify that the mutant was *S. pneumoniae* and not a contaminating organism. For the Δ*ply* mutant primers were used for the SPN23F13160 gene for this verification step, as detailed in chapter 2, section 2.3.4.5.

Appendix II

II.I Copyright licences

Figure 1.1 - Licence number: 4611400905095

Figure 1.5 - Licence number: 4611430813917

Figure 1.11 - Licence number: 4611440461076

Figure 1.14 - Licence number: 4611820591121

Figure 1.15 - Licence number: 4612540600795

Figure 1.16 – Licence number: 4725390692223

All other figures obtained from journal articles are available to reproduce in a thesis or dissertation without permission from the publisher, provided the author and publisher are credited in the figure legend.

Any figure that does not credit a journal article has been created for this thesis.

Bibliography

Abdelnour, A. et al., 1993. The accessory gene regulator (*agr*) controls *Staphylococcus aureus* virulence in a murine arthritis model. *Infection and Immunity*, 61(9), pp.3879-3885.

Alam, M.T. et al., 2014. Dissecting vancomycin-intermediate resistance in *Staphylococcus aureus* using genome-wide association. *Genome Biology and Evolution*, 6(5), pp.1174-85.

Alloing, G., de Philip, P. & Claverys, J.P., 1994. Three highly homologous membrane-bound lipoproteins participate in oligopeptide transport by the Ami system of the Gram-positive *Streptococcus pneumoniae*. *Journal of Molecular Biology*, 241(1), pp.44-58.

Anderton, J.M. et al., 2007. E-cadherin is a receptor for the common protein pneumococcal surface adhesin A (PsaA) of *Streptococcus pneumoniae*. *Microbial Pathogenesis*, 42(5-6), pp.225-236.

Andre, G.O. et al., 2017. Role of *Streptococcus pneumoniae* proteins in evasion of complement-mediated immunity. *Frontiers in Microbiology*, 8:224.

Antoniou, A.C. et al., 2010. A locus on 19p13 modifies risk of breast cancer in *BRCA1* mutation carriers and is associated with hormone receptor-negative breast cancer in the general population. *Nature Genetics*, 42(10), pp.885-892.

Archer, N.K. et al., 2011. *Staphylococcus aureus* biofilms: Properties, regulation and roles in human disease. *Virulence*, 2(5), pp.445-459.

Armstrong, R.A., 2014. When to use the Bonferroni correction. *Opthalmic & Physiological Optics*, 34(5), pp.502-508.

Arvidson, S. & Tegmark, K., 2001. Regulation of virulence determinants in *Staphylococcus aureus*. *International Journal of Medical Microbiology*, 291(2), pp.159-170.

Avery, O.T. & Dubos, R., 1931. The protective action of a specific enzyme against type III pneumococcus infection in mice. *The Journal of Experimental Medicine*, 54(1), pp.73-89.

Avery O.T., MacLeod C.M. & McCarty M., 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a deoxyribonucleic acid fraction isolated from pneumococcus type III. *The Journal of Experimental Medicine*, 79(2), pp.137–158.

Backes, C. et al., 2014. Systematic permutation testing in GWAS pathway analyses: identification of genetic networks in dilated cardiomyopathy and ulcerative colitis. *BMC Genomics*, 15:622.

Badiou, C. et al., 2010. Rapid detection of *Staphylococcus aureus* Panton-Valentine leukocidin in clinical specimens by enzyme-linked immunosorbent assay and immunochromatographic tests. *Journal of Clinical Microbiology*, 48(4), pp.1384-1390.

Bae, T. et al., 2008. Generating a collection of insertion mutations in the *Staphylococcus aureus* genome using bursa aurealis. In: A.L. Osterman & S.Y. Gerdes eds., 2008. *Methods in Molecular Biology* V*olume 416: Microbial Gene Essentiality*. New Jersey: Humana Press, pp.103–116.

Bagnoli, F. et al., 2008. A second pilus type in *Streptococcus pneumoniae* is prevalent in emerging serotypes and mediates adhesion to host cells. *Journal of Bacteriology*, 190(15), pp.5480-5492.

Balachandran, P. et al., 2001. The autolytic enzyme LytA of *Streptococcus pneumoniae* is not responsible for releasing pneumolysin. *Journal of Bacteriology*, 183(10), pp.3108-3116.

Barocchi, M.A. et al., 2005. A pneumococcal pilus influences virulence and host inflammatory responses. *PNAS*, 103(8), pp.2857-2862.

Bartilson, M. et al., 2001. Differential fluorescence induction reveals *Streptococcus pneumoniae* loci regulated by competence stimulatory peptide. *Molecular Microbiology*, 39(1), pp.126-135.

Bender, M.H., Cartee, R.T. & Yother, J., 2003. Positive correlation between tyrosine phosphorylation of CpsD and capsular polysaccharide production in *Streptococcus pneumoniae*. *Journal of Bacteriology*, 185(20), pp.6057-6066.

Benjamini, Y. & Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1), pp.289-300.

- For 1990 paper, see reference 'Hochberg & Benjamini' below.

Bergmann, S. & Hammerschmidt, S., 2006. Versatility of pneumococcal surface proteins. *Microbiology*, 152, pp.295-303.

Bergsbaken, T., Fink, S.L. & Cookson, B.T., 2009. Pyroptosis: host cell death and inflammation. *Nature Reviews Microbiology*, 7(2), pp.99-109.

Berry, A.M. et al., 1989. Contribution of autolysin to virulence of *Streptococcus pneumoniae*. *Infection and Immunity*, 57(8), pp.2324-2330.

Berry, A.M. & Paton, J.C., 1996. Sequence heterogeneity of PsaA, a 37-kilodalton putative adhesin essential for virulence of *Streptococcus pneumoniae*. *Infection and Immunity*, 64(12), pp.5255-5262.

Berube, B.J. & Wardenburg, J.B., 2013. *Staphylococcus aureus* α-toxin: nearly a century of intrigue. *Toxins*, 5(6), pp.1140-1166.

Bethe, G. et al., 2001. The cell wall-associated serine protease PrtA: a highly conserved virulence factor of *Streptococcus pneumoniae*. *FEMS Microbiology Letters*, 205(1), pp.99-104.

Bewley, M. et al., 2014. Pneumolysin activates macrophage lysosomal membrane permeabilization and executes apoptosis by distinct mechanisms without membrane pore formation. *mBio*, 5(5), e01710-14.

Bhakdi, S. & Tranum-Jensen, J., 1991. Alpha-toxin of *Staphylococcus aureus*. *Microbiological Reviews*, 55(4), pp.733-751.

Biasini, M. et al., 2014. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research*, 42(W1), pp.W252-W258.

Blaskovich, M.A.T., Butler, M.S. & Cooper, M.A., 2017. Polishing the tarnished silver bullet: the quest for new antibiotics. *Essays in Biochemistry*, 61, pp.103-114.

Boc, A. & Makarenkov, V., 2011. Towards an accurate identification of mosaic genes and partial horizontal gene transfers. *Nucleic Acids Research*, 39(21), e144.

Bogaert, D., de Groot, R. & Hermans, PWM., 2004. *Streptococcus pneumoniae* colonisation: the key to pneumococcal disease. *The Lancet Infectious Diseases*, 4(3), pp.144-154.

Bogaert D. et al., 2004. Colonisation by *Streptococcus pneumoniae* and *Staphylococcus aureus* in healthy children. *The Lancet*, 363(9424), pp.1871-1872.

Bokarewa, M.I., Jin, T. & Tarkowski, A., 2006. *Staphylococcus aureus*: staphylokinase. *The International Journal of Biochemistry and Cell Biology*, 38(4), pp.504-509.

Boles, B.R. & Horswill, A.R., 2008. *agr*-mediated dispersal of *Staphylococcus aureus* biofilms. *PLoS Pathogens*, 4(4), e1000052.

Boucher, H.W. & Corey, G.R., 2008. Epidemiology of methicillin-resistant *Staphylococcus aureus*. *Clinical Infectious Diseases*, 46(Suppl. 5), pp.S344-S349.

Boulnois, G.J., 1992. Pneumococcal proteins and the pathogenesis of disease caused by *Streptococcus pneumoniae*. *Journal of General Microbiology*, 138(2), pp.249-259.

Brittan, J.L. et al., 2012. Pneumococcal neuraminidase A: an essential upper airway colonisation factor for *Streptococcus pneumoniae*. *Molecular Oral Microbiology*, 27(4), pp.270-283.

Bronner, S. Monteil, H. & Prévost, G., 2004. Regulation of virulence determinants in *Staphylococcus aureus*: complexity and applications. *FEMS Microbiology Reviews*, 28(2), pp.183-200.

Brown, J.S., Gilliland, S.M. & Holden, D.W., 2001. A *Streptococcus pneumoniae* pathogenicity island encoding an ABC transporter involved in iron uptake and virulence. *Molecular Microbiology*, 40(3), pp.572-585.

Brueggemann, A.B. et al., 2003. Clonal relationships between invasive and carriage *Streptococcus pneumoniae* and serotype- and clone-specific differences in invasive disease potential. *The Journal of Infectious Diseases*, 187(9), pp.1424-1432.

Brynildsrud, O. et al., 2016. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biology*, 17:238.

Brzyski, D. et al., 2017. Controlling the rate of GWAS false discoveries. *Genetics*, 205(1), pp.61-75.

Bush, W.S. & Moore, J.H., 2012. Chapter 11: genome-wide association studies. *PLoS Computational Biology*, 8(12), e1002822.

Cantor, R.M., Lange, K. & Sinsheimer, J.S., 2010. Prioritising GWAS results: a review of statistical methods and recommendations for their application. *The American Journal of Human Genetics*, 86(1), pp.6-22.

Canvin, J.R. et al., 1995. The role of pneumolysin and autolysin in the pathology of pneumonia and septicaemia in mice infected with a type 2 pneumococcus. *The Journal of Infectious Diseases*, 172(1), pp.119-123.

Casadevall, A. & Pirofski, L.A., 1999. Host-pathogen interactions: redefining the basic concepts of virulence and pathogenicity. *Infection and Immunity*, 67(8), pp.3703-3713.

Centers for Disease Control and Prevention (CDC), 1999. Four pediatric deaths from community-acquired methicillin-resistant *Staphylococcus aureus* – Minnesota and North Dakota, 1997-1999. *Morbidity and Mortality Weekly Report*, 48(32), pp.707-710.

Chan, P.F. et al., 1998. The *Staphylococcus aureus* alternative sigma factor σ^B controls the environmental stress response but not starvation survival or pathogenicity in a mouse abscess model. *Journal of Bacteriology*, 180(23), pp.6082-6089.

Chan, P.F. & Foster, S.J., 1998. Role of SarA in virulence determinant production and environmental signal transduction in *Staphylococcus aureus*. *Journal of Bacteriology*, 180(23), pp.6232-6241.

Chaston, J.M., Newell, P.D. & Douglas, A.E., 2014. Metagenome-wide association of microbial determinants of host phenotype in *Drosophila melanogaster*. *mBio*, 5(5), e01631-14.

Chen, L., Ge, X. & Xu, P., 2015. Identifying essential *Streptococcus sanguinis* genes using genome-wide deletion mutation. *Methods in Molecular Biology*, 1279, pp.15-23.

Chen, P.E. & Shapiro, B.J., 2015. The advent of genome-wide association studies for bacteria. *Current Opinion in Microbiology*, 25, pp.17-24.

Cheung, A.L., Eberhardt, K. & Heinrichs, J.H., 1997. Regulation of protein A synthesis by the *sar* and *agr* loci of *Staphylococcus aureus*. *Infection and Immunity*, 65(6), pp.2243-2249.

Chewapreecha, C. et al., 2014. Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. *PLoS Genetics*, 10(8), e1004547.

Chi, Y.C. et al., 2016. *Streptococcus pneumoniae* IgA1 protease: A metalloprotease that can catalyze in a split manner *in vitro*. *Protein Science*, 26(3), pp.600-610.

Chien, Y.W. et al., 2013. Density interactions between *Streptococcus pneumoniae*, *Haemophilus influenzae* and *Staphylococcus aureus* in the nasopharynx of young Peruvian children. *The Pediatric Infectious Disease Journal*, 32(1), pp.72-77.

Clauditz, A. et al., 2006. Staphyloxanthin plays a role in the fitness of *Staphylococcus aureus* and its ability to cope with oxidative stress. *Infection and Immunity*, 74(8), pp.4950-4953.

Claverys, J.P., Grossoird, B., & Alloing, G., 2000. Is the Ami-AliA/B oligopeptide permease of *Streptococcus pneumoniae* involved in sensing environmental conditions? *Research in Microbiology*, 151(6), pp.457-463.

Claverys, J.P. & Håvarstein, L.S., 2007. Cannibalism and fratricide: mechanisms and raisons d'être. *Nature Reviews Microbiology*, 5(3), pp.219-229.

Claverys, J.P., Prudhomme, M. & Martin, B., 2006. Induction of competence regulons as a general response to stress in Gram-positive bacteria. *Annual Review of Microbiology*, 60, pp.451-475.

Cockeran, R. et al., 2002. Pneumolysin activates the synthesis and release of interleukin-8 by human neutrophils *in vitro*. *The Journal of Infectious Diseases*, 186(4), pp.562-565.

Coleman, G., 1985. A comparison of the patterns of extracellular proteins produced by the high alpha-toxin-secreting organism *Staphylococcus aureus* (Wood 46) during aerobic and anaerobic growth. *Journal of General Microbiology*, 131(2), pp.405-408.

Collins, C. & Didelot, X., 2018. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLOS Computational Biology*, 14(2), e1005958.

Costello, R. et al., 2016. Characterisation of isolates of *Streptococcus pneumoniae* in patients with chronic obstructive pulmonary disease – is there a role for the pneumococcal conjugate vaccine (pcv13)? *American Journal of Respiratory and Critical Care Medicine*, 193:A5188.

Costerton, J.W., Stewart, P.S. & Greenberg, E.P., 1999. Bacterial biofilms: a common cause of persistent infections. *Science*, 284(5418), pp.1318-1322.

Croucher, N.J. et al., 2009. Role of conjugative elements in the evolution of the multidrug-resistant pandemic clone *Streptococcus pneumoniae*^{SPN23F} ST81. *Journal of Bacteriology*, 191(5), pp.1480-1489.

Croucher, N.J. et al., 2011. Rapid pneumococcal evolution in response to clinical interventions. *Science*, 331(6016), pp.430-434.

Cundell, D.R. et al., 1995. Peptide permeases from *Streptococcus pneumoniae* affect adherence to eucaryotic cells. *Infection and Immunity*, 63(7), pp.2493-2498.

Cundell, D., Masure, H.R. & Tuomanen, E.I., 1995. The molecular basis of pneumococcal infection: A hypothesis. *Clinical Infectious Diseases*, 21(3), pp.204-212.

Das, S. et al., 2016. Natural mutations in a *Staphylococcus aureus* virulence regulator attenuate cytotoxicity but permit bacteraemia and abscess formation. *PNAS*, 113(22), pp.E3101-E3110.

Davenport, E.R. et al., 2015. Genome-wide association studies of the human gut microbiota. *PLoS One*, 10(11), e0140301.

David, M.Z. et al., 2011. Methicillin-susceptible *Staphylococcus aureus* as a predominantly healthcare-associated pathogen: a possible reversal of roles? *PLoS One*, 6(4), e18217.

Deng, X. et al., 2016. Whole-genome sequencing reveals the origin and rapid evolution of an emerging outbreak strain of *Streptococcus pneumoniae* 12F. *Clinical Infectious Diseases*, 62(9), pp.1126-1132.

de Stoppelaar, S.F. et al., 2013. *Streptococcus pneumoniae* serine protease HtrA, but not SFP or PrtA, is a major virulence factor in pneumonia. *PLoS One*, 8(11), e80062.

Devaux, C.A., Mezouar, S. & Mege, J.L., 2019. The E-cadherin cleavage associated to pathogenic bacteria infections can favour bacterial invasion and transmigration, dysregulation of the immune response and cancer induction in humans. *Frontiers in Microbiology*, 10:2598.

Dickerson, R.E., Timkovich, R. & Almassy, R.J., 1976. The cytochrome fold and the evolution of bacterial energy metabolism. *Journal of Molecular Biology*, 100(4), pp.473-491.

Dinges, M.M., Orwin, P.M. & Schlievert, P.M., 2000. Exotoxins of *Staphylococcus aureus*. *Clinical Microbiology Reviews*, 13(1), pp.16-34.

Dintilhac, A. et al., 1997. Competence and virulence of *Streptococcus pneumoniae*: Adc and PsaA mutants exhibit a requirement for Zn and Mn resulting from inactivation of putative ABC metal permeases. *Molecular Microbiology*, 25(4), pp.727-739.

Dobson, A., 2004. Population dynamics of pathogens with multiple host species. *The American Naturalist*, 164(Suppl. 5), pp.S64-S78.

Earle, S.G. et al., 2016. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nature Microbiology*, 1:16041.

Eidhin, D.N. et al., 1998. Clumpin factor B (ClfB), a new surface located fibrinogen-binding adhesin of *Staphylococcus aureus*. *Molecular Microbiology*, 30(2), pp.245-257.

Elasri, M.O. et al., 2002. *Staphylococcus aureus* collagen adhesin contributes to the pathogenesis of osteomyelitis. *Bone*, 30(1), pp.275-280.

Fang, R. et al., 2011. Critical roles of ASC inflammasomes in caspase-1 activation and host innate resistance to *Streptococcus pneumoniae* infection. *The Journal of Immunology*, 187(9), pp.4890-4899.

Farhat, M.R., et al., 2013. Genomic analysis identifies targets of convergent positive selection in drug resistant *Mycobacterium tuberculosis*. *Nature Genetics*, 45(10), pp.1183-1189.

Faure, G. et al., 2019. CRISPR-Cas in mobile genetic elements: counter-defence and beyond. *Nature Reviews Microbiology*, 17(8), pp.513-525

Fey, P.D. et al., 2013. A genetic resource for rapid and comprehensive phenotype screening of nonessential *Staphylococcus aureus* genes. *mBio*, 4(1), e00537-12.

Finlay, B.B. & Hancock, R.E.W., 2004. Can innate immunity be enhanced to treat microbial infections? *Nature Reviews Microbiology*, 2(6), pp.497-504.

Foster. T.J. & Höök, M., 1998. Surface protein adhesins of *Staphylococcus aureus*. *Trends in Microbiology*, 6(12), pp.484-488.

Foster, T.J., 2005. Immune evasion by staphylococci. *Nature Reviews Microbiology*, 3(12), pp.948-958.

Foster, T.J. et al., 2014. Adhesion, invasion and evasion: the many functions of the surface proteins of *Staphylococcus aureus*. *Nature Reviews Microbiology*, 12(1), pp.49-62.

Ganesh, V.K. et al., 2008. A structural model of the *Staphylococcus aureus* ClfA-fibrinogen interaction opens new avenues for the design of anti-staphylococcal therapeutics. *PLoS Pathogens*, 4(11), e1000226.

Ganesh, V.K. et al., 2011. Structural and biochemical characterization of *Staphylococcus aureus* clumping factor B/ligand interactions. *Journal of Biological Chemistry*, 286(29), pp.25963-25972.

Gardy, J., Loman, N.J. & Rambaut, A., 2015. Real-time digital pathogen surveillance – the time is now. *Genome Biology*, 16:155.

Ge, X. & Xu, P., 2012. Genome-wide deletions in *Streptococcus sanguinis* by high-throughput PCR. *Journal of Visualised Experiments*, 69, e4356.

Geiger, T. et al., 2008. The virulence regulator Sae of *Staphylococcus aureus*: promoter activities and response to phagocytosis-related signals. *Journal of Bacteriology*, 190(10), pp.3419-3438.

Gemmell, C. et al., 1991. Susceptibility to opsonophagocytosis of protein A, α-haemolysin and β-toxin deficient mutants of *Staphylococcus aureus* isolated by allele-replacement. *Zentralblatt für Bakteriologie*, 21(Suppl.), pp.273-277.

Georgellis, D., Kwon, O. & Lin, E.C.C., 2001. Quinones as the redox signal for the Arc two-component system of bacteria. *Science*, 292(5525), pp.2314-2316.

Gillaspy, A.F. et al., 1995. Role of the accessory gene regulator (*agr*) in pathogenesis of staphylococcal osteomyelitis. *Infection and Immunity*, 63(9), pp.3373-3380.

Gillet, Y. et al., 2002. Association between *Staphylococcus aureus* strains carrying gene for Panton-Valentine leukocidin and highly lethal necrotising pneumonia in young immunocompetent patients. *The Lancet*, 359(9308), pp.753-759.

Gómez, M.I. et al., 2004. *Staphylococcus aureus* protein A induces airway epithelial inflammatory responses by activating TNFR1. *Nature Medicine*, 10(8), pp.842-848.

Goodyear, C.S. & Silverman, G.J., 2003. Death by a B cell superantigen: *in vivo* V_H-targeted apoptotic supraclonal B cell deletion by a staphylococcal toxin. *The Journal of Experimental Medicine*, 197(9), pp.1125-1139.

Gordon, R.J. & Lowy, F.D., 2008. Pathogenesis of methicillin-resistant *Staphylococcus aureus* infection. *Clinical Infectious Diseases*, 46(5), pp.350-359.

Gotoh, Y. et al., 2010. Two-component signal transduction as potential drug targets in pathogenic bacteria. *Current Opinion in Microbiology*, 13(2), pp.232-239.

Gouaux, E., Hobaugh, M. & Song, L., 1997. α-haemolysin, γ-haemolysin, and leukocidin from *Staphylococcus aureus*: Distant in sequence but similar in structure. *Protein Science*, 6(12), pp.2631-2635.

Gould, I.M., 2006. Costs of hospital-acquired methicillin-resistant *Staphylococcus aureus* (MRSA) and its control. *International Journal of Antimicrobial Agents*, 28(5), pp.379-384.

Graves, S.F. et al., 2010. Relative contribution of Panton-Valentine leukocidin to PMN plasma membrane permeability and lysis caused by USA300 and USA400 culture supernatants. *Microbes & Infection*, 12(6), pp.446-456.

Gross, M., 2013. Antibiotics in crisis. *Current Biology*, 23(24), pp.R1063-R1065.

Guérillot, R. et al., 2014. The diversity of prokaryotic DDE transposases of the mutator superfamily, insertion specificity, and association with conjugation machineries. *Genome Biology and Evolution*, 6(2), pp.260-272.

Guiral, S. et al., 2005. Competence-programmed predation of noncompetent cells in the human pathogen *Streptococcus pneumoniae*: Genetic requirements. *PNAS*, 102(24), pp.8710-8715.

Guo, Y.F. et al., 2009. A new permutation strategy of pathway-based approach for genome-wide association study. *BMC Bioinformatics*, 10:429.

Gupta, R.K., Luong, T.T. & Lee, C.Y., 2015. RNAIII of the *Staphylococcus aureus agr* system activates global regulator MgrA by stabilizing mRNA. *PNAS*, 112(45), pp.14036-14041.

Hadfield, J. et al., 2017. Phandango: an interactive viewer for bacterial population genomics. *Bioinformatics*, 34(2), pp.292-293.

Hammerschmidt, S. et al., 2005. Illustration of pneumococcal polysaccharide capsule during adherence and invasion of epithelial cells. *Infection and Immunity*, 73(8), pp.4653-4667.

Haque, M. et al., 2018. Healthcare-associated infections – an overview. *Infection and Drug Resistance*, 11, pp.2321-2333.

Harris, S.R. et al., 2010. Evolution of MRSA during hospital transmission and intercontinental spread. *Science*, 327(5964), pp.469-474.

Hava, D.L. & Camilli, A., 2002. Large-scale identification of serotype 4 *Streptococcus pneumoniae* virulence factors. *Molecular Microbiology*, 45(5), pp.1389-1406.

Hendley, J.O. et al., 1975. Spread of *Streptococcus pneumoniae* in families: 1. Carriage rates and distribution of types. *The Journal of Infectious Diseases*, 132(1), pp.55-61.

Henrichsen, J., 1995. Six newly recognised types of *Streptococcus pneumoniae*. *Journal of Clinical Microbiology*, 33(10), pp.2759-62.

Heilmann, C., 2011. Adhesion mechanisms of staphylococci. In: D. Linke & A. Goldman eds., 2011. *Bacterial adhesion: Advances in Experimental Medicine and Biology Volume 715*. Dordrecht: Springer, pp.105-123.

Henriques-Normark, B. & Normark, S., 2010. Commensal pathogens, with a focus on *Streptococcus pneumoniae*, and interactions with the human host. *Experimental Cell Research*, 316(8), pp.1408-1414.

Hienz, S.A. et al., 1996. Collagen binding of *Staphylococcus aureus* is a virulence factor in experimental endocarditis. *The Journal of Infectious Diseases*, 174(1), pp.83-88.

Higgins, J. et al., 2006. Clumping factor A of *Staphylococcus aureus* inhibits phagocytosis by human polymorphonuclear leucocytes. *FEMS Microbiology Letters*, 258(2), pp.290-296.

Hill, J. et al., 1992. Demonstration by FTIR that the *bo*-type ubiquinol oxidase of *Escherichia coli* contains a heme-copper binuclear center similar to that in cytochrome *c* oxidase and that proper assembly of the binuclear center requires the *cyoE* gene product. *Biochemistry*, 31(46), pp.11435-11440.

Hirst, R.A. et al., 2004. The role of pneumolysin in pneumococcal pneumonia and meningitis. *Clinical & Experimental Immunology*, 138(2), pp.195-201.

Hirst, R.A. et al., 2008. *Streptococcus pneumoniae* deficient in pneumolysin or autolysin has reduced virulence in meningitis. *The Journal of Infectious Diseases*, 197(5), pp.744-751.

Hochberg, Y. & Benjamini, Y., 1990. More powerful procedures for multiple significance testing. *Statistics in Medicine*, 9(7), pp.811-818.

Hoegen, T. et al., 2011. The NLRP3 inflammasome contributes to brain injury in pneumococcal meningitis and is activated through ATP-dependent lysosomal cathepsin B release. *The Journal of Immunology*, 187(10), pp.5440-5451.

Holden, M.T.G. et al., 2010. Genome sequence of a recently emerged, highly transmissible, multi-antibiotic- and antiseptic-resistant variant of methicillin-resistant *Staphylococcus aureus*, sequence type 239 (TW). *Journal of Bacteriology*, 192(3), pp.888-892.

Hotomi, M. et al., 2016. Pneumolysin plays a key role at the initial step of establishing pneumococcal nasal colonization. *Folia Microbiologica*, 61(5), pp.375-383.

Houldsworth, S., Andrew, P.W. & Mitchell, T.J., 1994. Pneumolysin stimulates production of tumouor necrosis factor alpha and interleukin-1β by human mononuclear phagocytes. *Infection and Immunity*, 62(4), pp.1501-1503.

Howard, L.V. & Gooder, H., 1974. Specificity of the autolysin of *Streptococcus* (*Diplococcus*) *pneumoniae*. *Journal of Bacteriology*, 117(2), pp.796-804.

Hsu, C.F. et al., 2018. PrtA immunization fails to protect against pulmonary and invasive infection by *Streptococcus pneumoniae*. *Respiratory Research*, 19:187.

Hupp, S. et al., 2019. Pneumolysin and the bacterial capsule of *Streptococcus pneumoniae* cooperatively inhibit taxis and motility of microglia. *Journal of Neuroinflammation*, 16:105.

Hussain, M. et al., 2005. A longitudinal household study of *Streptococcus pneumoniae* nasopharyngeal carriage in a UK setting. *Epidemiology & Infection*, 133(5), pp.891-898.

Hyams, C. et al., 2010. The *Streptococcus pneumoniae* capsule inhibits complement activity and neutrophil phagocytosis by multiple mechanisms. *Infection and Immunity*, 78(2), pp.704-715.

Ibrahim, Y.M. et al., 2004. Role of HtrA in the virulence and competence of *Streptococcus pneumoniae*. *Infection and Immunity*, 72(6), pp.3584-3591.

Inoshima, I. et al., 2011. A *Staphylococcus aureus* pore-forming toxin subverts the activity of ADAM10 to cause lethal infection in mice. *Nature Medicine*, 17(10), pp.1310-1314.

Ioachimescu, O.C., Ioachimescu, A.G. & Iannini, P.B., 2004. Severity scoring in community-acquired pneumonia caused by *Streptococcus pneumoniae*: a 5-year experience. *International Journal of Antimicrobial Agents*, 24(5), pp.485-490.

Janzon, L & Arvidson, S., 1990. The role of the δ-haemolysin gene (*hld*) in the regulation of virulence genes by the accessory gene regulator (*agr*) in *Staphylococcus aureus*. *The EMBO Journal*, 9(5), pp.1391-1399.

Jenkinson, H.F., 1997. Streptococcal adhesion and colonization. *Critical Reviews in Oral Biology and Medicine*, 8(2), pp.175-200.

Ji, G., Beavis, R. & Novick, R.P., 1997. Bacterial interference caused by autoinducing peptide variants. *Science*, 276(5321), pp.2027-2030.

Johnson, A.P. et al., 2012. Mandatory surveillance of methicillin-resistant *Staphylococcus aureus* (MRSA) bacteraemia in England: the first 10 years. *Journal of Antimicrobial Chemotherapy*, 67(4), pp.802-809.

Josefsson, E. et al., 2001. Protection against experimental *Staphylococcus aureus* arthritis by vaccination with clumping factor A, a novel virulence determinant. *The Journal of Infectious Diseases*, 184(12), pp.1572-1580.

Kadioglu, A. et al., 2008. The role of *Streptococcus pneumoniae* virulence factors in host respiratory colonization and disease. *Nature Reviews Microbiology*, 6(4), pp.288-301.

Karakawa, W.W. et al., 1988. Capsular antibodies induce type-specific phagocytosis of capsulated *Staphylococcus aureus* by human polymorphonuclear leukocytes. *Infection and Immunity*, 56(5), pp.1090-1095.

Kennedy, M.C. et al., 1983. The role of iron in the activation-inactivation of aconitase. *The Journal of Biological Chemistry*, 258(18), pp.11098-11105.

Kim, J.O. & Weiser, J.N., 1998. Association of intrastain phase variation in quantity of capsular polysaccharide and teichoic acid with the virulence of *Streptococcus pneumoniae*. *The Journal of Infectious Diseases*, 177(2), pp.368-377.

King, S.J. et al., 2004. Phase variable desialylation of host proteins that bind to *Streptococcus pneumoniae in vivo* and protect the airway. *Molecular Microbiology*, 54(1), pp.159-171.

King, S.J., Hippe, K.R. & Weiser, J.N., 2006. Deglycosylation of human glycoconjugates by the sequential activities of exoglycosidases expressed by *Streptococcus pneumoniae*. *Molecular Microbiology*, 59(3), pp.961-974.

Kluytmans, J. et al., 1997. Nasal carriage of *Staphylococcus aureus*: Epidemiology, underlying mechanisms, and associated risks. *Clinical Microbiology Reviews*, 10(3), pp.505-520.

Knutsen, E. et al., 2006. BOX elements modulate gene expression in *Streptococcus pneumoniae*: Impact on the fine-tuning of competence development. *Journal of Bacteriology*, 188(23), pp.8307-8312.

Ko, K.S., Baek, J.Y. & Song, J.H., 2012. Multidrug-resistant *Streptococcus pneumoniae* serotype 6D clones in South Korea. *Journal of Clinical Microbiology*, 50(3), pp.818-822.

Koenig, R.L. et al., 2004. *Staphylococcus aureus* AgrA binding to the RNAIII-*agr* regulatory region. *Journal of Bacteriology*, 186(22), pp.7549-7555.

Kohler, C. et al., 2003. Physiological characterisation of a heme-deficient mutant of *Staphylococcus aureus* by a proteomic approach. *Journal of Bacteriology*, 185(23), pp.6928-6937.

Köck, R. et al., 2010. Methicillin-resistant *Staphylococcus aureus* (MRSA): burden of disease and control challenges in Europe. *Euro Surveillance*, 15(41), e19688.

Kraft, P. & Haiman, C.A., 2010. GWAS identifies a common breast cancer risk allele among *BRCA1* carriers. *Nature Genetics*, 42(10), pp.819-820.

Kyme, P. et al., 2012. C/EBPε mediates nicotinamide-enhanced clearance of *Staphylococcus aureus* in mice. *Journal of Clinical Investigation*, 122(9), pp.3316-3329.

Laabei, M. et al., 2014. Predicting the virulence of MRSA from its genome sequence. *Genome Research*, 24(5), pp.839-849.

Laabei, M. et al., 2015. Evolutionary trade-offs underlie the multifaceted virulence of *Staphylococcus aureus*. *PLoS Biology*, 13(9), e1002229.

Lau, G.W. et al., 2001. A functional genomic analysis of type 3 *Streptococcus pneumoniae* virulence. *Molecular Microbiology*, 40(3), pp.555-571.

Lee, D.S. et al., 2009. Comparative genome-scale metabolic reconstruction and flux balance analysis of multiple *Staphylococcus aureus* genomes identify novel antimicrobial drug targets. *Journal of Bacteriology*, 191(12), pp.4015-4024.

Lee, L.Y.L. et al., 2004. Identification and characterization of the C3 binding domain of the *Staphylococcus aureus* extracellular fibrinogen-binding protein (Efb). *The Journal of Biological Chemistry*, 279(49), pp.50710-50716.

Lees, J.A. et al., 2016. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nature Communications*, 7:12797.

Levy, S.B., 1998. The challenge of antibiotic resistance. *Scientific American*, 278(3), pp.46-53.

Lin, J., Zhu, L. & Lau, G.W., 2016. Disentangling competence for genetic transformation and virulence in *Streptococcus pneumoniae*. *Current Genetics*, 62(1), pp.97-103.

Lina, G. et al., 1998. Transmembrane topology and histidine protein kinase activity of AgrC, the *agr* signal receptor in *Staphylococcus aureus*. *Molecular Microbiology*, 28(3), pp.655-662.

Littman, M. et al., 2009. *Streptococcus pneumoniae* evades human dendritic cell surveillance by pneumolysin expression. *EMBO Molecular Medicine*, 1(4), pp.211-222.

Liu, G.Y. et al., 2005. *Staphylococcus aureus* golden pigment impairs neutrophil killing and promotes virulence through its antioxidant activity. *Journal of Experimental Medicine*, 202(2), pp.209-215.

Liu, T.T. et al., 2012. Cardiac device infection due to *Streptococcus pneumoniae*. *The Canadian Journal of Infectious Diseases and Medical Microbiology*, 23(3), pp.135-136.

Louwen, R. et al., 2014. The role of CRISPR-Cas systems in virulence of pathogenic bacteria. *Microbiology and Molecular Biology Reviews*, 78(1), pp.74-88.

Lowy, F., 1998. *Staphylococcus aureus* infections. *The New England Journal of Medicine*, 339(8), pp.520-532.

Luong, T.T. & Lee, C.Y., 2002. Overproduction of type 8 capsular polysaccharide augments *Staphylococcus aureus* virulence. *Infection and Immunity*, 70(7), pp.3389-3395.

Magee, A.D. & Yother, J., 2001. Requirement for capsule in colonisation by *Streptococcus pneumoniae*. *Infection and Immunity*, 69(6), pp.3755-3761.

Malley, R. et al., 2003. Recognition of pneumolysin by Toll-like receptor 4 confers resistance to pneumococcal infection. *PNAS*, 100(4), pp.1966-1971.

Malpica, R. et al., 2004. Identification of a quinone-sensitive redox switch in the ArcB sensor kinase. *PNAS*, 101(36), pp.13318-13323.

Mamo, W., Bodén, M. & Flock, J.I., 1994. Vaccination with *Staphylococcus aureus* fibrinogen binding proteins (FgBPs) reduces colonisation of *S. aureus* in a mouse mastitis model. *FEMS Immunology and Medical Microbiology*, 10(1), pp.47-54.

Marrie, T.J., Nelligan, J. & Costerton, J.W., 1982. A scanning and transmission electron microscopy study of an infected endocardial pacemaker lead. *Circulation*, 66(6), pp.1339-1341.

Martner, A. et al., 2009. *Streptococcus pneumoniae* autolysis prevents phagocytosis and production of phagocyte-activating cytokines. *Infection and Immunity*, 77(9), pp.3826-3837.

Massey, R.C. et al., 2001. Fibronectin-binding protein A of *Staphylococcus aureus* has multiple, substituting, binding regions that mediate adherence to fibronectin and invasion of endothelial cells. *Cellular Microbiology*, 3(12), pp.839-851.

Maury, M.M. et al., 2016. Uncovering *Listeria monocytogenes* hypervirulence by harnessing its biodiversity. *Nature Genetics*, 48(3), pp.308-313.

McAdow, M., Missiakas, D.M. & Schneewind, O., 2012. *Staphylococcus aureus* secretes coagulase and von Willebrand factor binding protein to modify the coagulation cascade and establish host infections. *Journal of Innate Immunity*, 4(2), pp.141-148.

McCullers, J.A. & English, B.K., 2008. Improving therapeutic strategies for secondary bacterial pneumonia following influenza. *Future Microbiology*, 3(4), pp.397-404.

McNamara, P.J. & Proctor, R.A., 2000. *Staphylococcus aureus* small colony variants, electron transport and persistent infections. *International Journal of Antimicrobial Agents*, 14(2), pp.117-233.

McNeela, E.A. et al., 2010. Pneumolysin activates the NLRP3 inflammasome and promotes proinflammatory cytokines independently of TLR4. *PLoS Pathogens*, 6(11), e1001191.

Mehr, S. & Wood, N., 2012. *Streptococcus pneumoniae* – a review of carriage, infection, serotype replacement and vaccination. *Paediatric Respiratory Reviews*, 13(4), pp.258-264.

Mellroth, P. et al., 2012. LytA, major autolysin of *Streptococcus pneumoniae*, requires access to nascent peptidoglycan. *The Journal of Biological Chemistry*, 287(14), pp.11018-11029.

Mitchell, T.J. et al., 1991. Complement activation and antibody binding by pneumolysin via a region of the toxin homologous to a human acute-phase protein. *Molecular Microbiology*, 5(8), pp.1883-1888.

Mitchell, T.J., 2000. Virulence factors and the pathogenesis of disease caused by *Streptococcus pneumoniae*. *Research in Microbiology*, 151(6), pp.413-419.

Mitchell, A.M. & Mitchell, T.J., 2010. *Streptococcus pneumoniae*: virulence factors and variation. *Clinical Microbiology and Infectious Disease*, 16(5), pp.411-418.

Morales, M. et al., 2015. Insights into the evolutionary relationships of LytA autolysin and Ply pneumolysin-like genes in *Streptococcus pneumoniae* and related streptococci. *Genome Biology and Evolution*, 7(9), pp.2747-2761.

Morfeldt, E. et al., 1988. Cloning of a chromosomal locus (*exp*) which regulates the expression of several exoprotein genes in *Staphylococcus aureus*. *Molecular and General Genetics*, 211(3), pp.435-440.

Morona, J.K., Morona, R. & Paton, J.C., 1997. Characterization of the locus encoding the *Streptococcus pneumoniae* type 19F capsular polysaccharide biosynthetic pathway. *Molecular Microbiology*, 23(4), pp.751-763.

Mosser, J.L. & Tomasz, A., 1970. Choline-containing teichoic acid as a structural component of pneumococcal cell wall and its role in sensitivity to lysis by an autolytic enzyme. *The Journal of Biological Chemistry*, 245(2), pp.287-298.

Mulcahy, M.E. et al., 2012. Nasal colonisation by *Staphylococcus aureus* depends upon clumping factor B binding to the squamous epithelial cell envelope protein loricrin. *PLoS Pathogens*, 8(12), e1003092.

Naber, C.K., 2009. *Staphylococcus aureus* bacteremia: Epidemiology, pathophysiology, and management strategies. *Clinical Infectious Diseases*, 48(Suppl. 4), pp.S231-S237.

Nakamura, H. et al., 1997. Assignment and functional roles of the *cyoABCDE* gene products required for the *Escherichia coli bo*-type quinol oxidase. *Journal of Biochemistry*, 122(2), pp.415-421.

Nau, R. & Eiffert, H., 2002. Modulation of release of proinflammatory bacterial compounds by antibacterials: Potential impact on course of inflammation and outcome in sepsis and meningitis. *Clinical Microbiology Reviews*, 15(1), pp.95-110.

Nau, R. & Eiffert, H., 2005. Minimising the release of proinflammatory and toxic bacterial products within the host: A promising approach to improve outcome in life-threatening infections. *FEMS Immunology and Medical Microbiology*, 44(1), pp.1-16.

Navarre, W.W. & Schneewind, O., 1994. Proteolytic cleavage and cell wall anchoring at the LPXTG motif of surface proteins in Gram-positive bacteria. *Molecular Microbiology*, 14(1), pp.115-121.

Neill, D.R., Mitchell, T.J. & Kadioglu, A., 2015. Pneumolysin. In: J. Brown, S. Hammerschmidt & Carlos Orihuela, eds. 2015. *Streptococcus pneumoniae. Molecular mechanisms of host-pathogen interactions*. USA: Academic Press. pp.257-275.

Nel, J.G. et al., 2016a. Pneumolysin mediates platelet activation *in vitro*. *Lung*, 194(4), pp.589-593.

Nel, J.G. et al., 2016b. Pneumolysin activates neutrophil extracellular trap formation. *Clinical and Experimental Immunology*, 184(3), pp.358-367.

Neu, H.C., 1992. The crisis in antibiotic resistance. *Science*, 257(5073), pp.1064-1073.

Niederman, M.S. & Zumla, A., 2016. Understanding community-acquired respiratory tract infections: new concepts of disease pathogenesis and new management strategies. *Current Opinion in Pulmonary Medicine*, 22(3), pp.193-195.

Nilsson, I.M. et al., 1997. The role of staphylococcal polysaccharide microcapsule expression in septicemia and septic arthritis. *Infection and Immunity*, 65(10), pp.4216-4221.

Nilsson, I.M. et al., 1998. Vaccination with a recombinant fragment of collagen adhesin provides protection against *Staphylococcus aureus*-mediated septic death. *The Journal of Clinical Investigation*, 101(12), pp.2640-2649.

Nozohoor, S. et al., 1998. Virulence factors of *Staphylococcus aureus* in the pathogenesis of endocarditis. A comparative study of clinical isolates. *Zentralblatt für Bakteriologie*, 287(4), pp.433-447.

Obert, C. et al., 2006. Identification of a candidate *Streptococcus pneumoniae* core genome and regions of diversity correlated with invasive pneumococcal disease. *Infection and Immunity*, 74(8), pp.4766-4777.

O'Brien, K.L. et al., 2009. Burden of disease caused by *Streptococcus pneumoniae* in children younger than 5 years: global estimates. *The Lancet*, 374(9693), pp.893-902.

O'Connell, D.P. et al., 1997. The fibrinogen-binding MSCRAMM (clumping factor) of *Staphylococcus aureus* has a Ca^{2+}-dependent inhibitory site. *The Journal of Biological Chemistry*, 273(12), pp.6821-6829.

Ogunniyi, A.D., Giammarinaro, P. & Paton, J.C., 2002. The genes encodingviruelnce-associated proteins and the capsule of *Streptococcus pneumoniae* are upregulated and differentially expressed *in vivo*. *Microbiology*, 148(Pt. 7), pp.2045-2053.

Okada, Y. et al., 2012. Common variants at *CDKAL1* and *KLF9* are associated with body mass index in east Asian populations. *Nature Genetics*, 44(3), pp.302-308.

O'Neill, E. et al., 2008. A novel *Staphylococcus aureus* biofilm phenotype mediated by the fibronectin-binding proteins, FnBPA and FnBPB. *Journal of Bacteriology*, 190(11), pp.3835-3850.

Osmasits, U. et al., 2014. Protter: interactive protein feature visualisation and integration with experimental proteomic data. *Bioinformatics*, 30(6), pp.884-886.

Otto, M. et al., 1998. Structure of the pheromone peptide of the *Staphylococcus epidermidis agr* system. *FEBS Letters*, 424(1-2), pp.89-94.

Otto, M., 2008. Staphylococcal biofilms. *Current Topics in Microbiology and Immunology*, 322, pp.207-228.

Otto, M., 2014. *Staphylococcus aureus* toxins. *Current Opinion in Microbiology*, 17, pp.32-37.

Ozaki, K. et al., 2002. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility with myocardial infarction. *Nature Genetics*, 32(4), pp.650-654.

Palma, M. et al., 1998. Multiple binding sites in the interaction between an extracellular fibrinogen-binding protein from *Staphylococcus aureus* and fibrinogen. *The Journal of Biological Chemistry*, 273(21), pp.13177-13181.

Palmqvist, N. et al., 2002. Protein A is a virulence factor in *Staphylococcus aureus* arthritis and septic death. *Microbial Pathogenesis*, 33(5), pp.239-249.

Park, B., Nizet, V. & Liu, G.Y., 2008. Role of *Staphylococcus aureus* catalase in niche competition against *Streptococcus pneumoniae*. *Journal of Bacteriology*, 190(7), pp.2275-2278.

Parker, D. et al., 2009. The NanA neuraminidase of *Streptococcus pneumoniae* is involved in biofilm formation. *Infection and Immunity*, 77(9), pp.3722-3730.

Patel, A.H. et al., 1987. Virulence of protein A-deficient and α-toxin-deficient mutants of *Staphylococcus aureus* isolated by allele replacement. *Infection and Immunity*, 55(12), pp.3103-3110.

Paton, J.C., Rowan-Kelly, B. & Ferrante, A., 1984. Activation of human complement by the pneumococcal toxin pneumolysin. *Infection and Immunity*, 43(3), pp.1085-1087.

Paton, J.C., 1996. The contribution of pneumolysin to the pathogenicity of *Streptococcus pneumoniae*. *Trends in Microbiology*, 4(3), pp.103-106.

Paton, J.C., 1998. Novel pneumococcal surface proteins: role in virulence and vaccine potential. *Trends in Microbiology*, 6(3), pp.85-87.

Patti, J.M. et al., 1994. The *Staphylococcus aureus* collagen adhesin is a virulence determinant in experimental septic arthritis. *Infection and Immunity*, 62(1), pp.152-161.

Peacock, S.J. et al., 1999. Bacterial fibronectin-binding proteins and endothelial cell surface fibronectin mediate adherence of *Staphylococcus aureus* to resting endothelial cells. *Microbiology*, 145(Pt. 12), pp.3477-3486.

Peacock, S.J. et al., 2002. Virulent combinations of adhesin and toxin genes in natural populations of *Staphylococcus aureus*. *Infection and Immunity*, 70(9), pp.4987-4996.

Perneger, T.V., 1998. What's wrong with Bonferroni adjustments. *BMJ*, 316:1236.

Perrenoud, A. & Sauer, U., 2005. Impact of global transcriptional regulation by ArcA, ArcB, Cra, Crp, Cya, Fnr and Mlc on glucose catabolism in *Escherichia coli*. *Journal of Bacteriology*, 187(9), pp.3171-3179.

Peterson, P.K. et al., 1977. Effect of protein A on staphylococcal opsonisation. *Infection and Immunity*, 15(3), pp.760-764.

Peterson, P.K. et al., 1978. Influence of encapsulation on staphylococcal opsonization and phagocytosis by human polymorphonuclear leukocytes. *Infection and Immunity*, 19(3), pp.943-949.

Polissi, A. et al., 1998. Large-scale identification of virulence genes from *Streptococcus pneumoniae*. *Infection and Immunity*, 66(12), pp.5620-5629.

Price, K.E. & Camilli, A., 2009. Pneumolysin localizes to the cell wall of *Streptococcus pneumoniae*. *Journal of Bacteriology*, 191(7), pp.2163-2168.

Priest, N. et al., 2012. From genotype to phenotype: can systems biology be used to predict *Staphylococcus aureus* virulence? *Nature Reviews Microbiology*, 10(11), pp.791-797.

Proctor, R.A. & Peters, G., 1998. Small colony variants in staphylococcal infections: diagnostic and therapeutic implications. *Clinical Infectious Diseases*, 27(3), pp.419-422.

Public Health England, 2018a. *Staphylococcus aureus (MRSA and MSSA) bacteraemia: mandatory surveillance 2017/18*. [pdf] London: PHE Publications. Available at: <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/724361/S_aureus_summary_2018.pdf> [Accessed 12 February 2019].

Public Health England, 2018b. *Pneumococcal disease: cases caused by strains covered by Prevenar13 vaccine*. [online] Public Health England: Immunisation Department. Available at: <https://www.gov.uk/government/publications/pneumococcal-disease-cases-caused-by-strains-covered-by-prevenar-13-vaccine/pneumococcal-disease-cases-caused-by-strains-covered-by-prevenar-13-vaccine> [Accessed 3rd June 2019].

Public Health England, 2018c. *Pneumococcal disease infections caused by serotypes not in Prevenar13 vaccine*. [online] Public Health England: Immunisation Department. Available at: <https://www.gov.uk/government/publications/pneumococcal-disease-caused-by-strains-not-covered-by-prevenar-13-vaccine/pneumococcal-disease-infections-caused-by-serotypes-not-in-prevenar-13-vaccine> [Accessed 3rd June 2019].

Quick, J. et al., 2016. Real-time, portable genome sequencing for Ebola surveillance. *Nature*, 530(7589), pp.228-232.

Rai, P. et al., 2016. Pneumococcal pneumolysin induces DNA damage and cell cycle arrest. *Scientific Reports*, 6:22972.

Rajam, G. et al., 2008. Pneumococcal surface adhesin A (PsaA): A review. *Critical Reviews in Microbiology*, 34(3-4), pp.131-142.

Rajasree, K., Fasim, A. & Gopal, B., 2016. Conformational features of the *Staphylococcus aureus* AgrA-promoter interactions rationalise quorum-sensing triggered gene expression. *Biochemistry and Biophysics Reports*, 6, pp.124.134.

Ratner, H.K. et al., 2019. Catalytically active Cas9 mediates transcriptional interference to facilitate bacterial virulence. *Molecular Cell*, 75(3), pp.498-510.

Read, T.D. & Massey, R.C., 2014. Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology. *Genome Medicine*, 6:109.

Recker, M. et al., 2017. Clonal differences in *Staphylococcus aureus* bacteraemia-associated mortality. *Nature Microbiology*, 2(10), pp.1381-1388.

Recsei, P. et al., 1986. Regulation of exoprotein gene expression in *Staphylococcus aureus* by *agr*. *Molecular and General Genetics*, 202(1), pp.58-61.

Regev-Yochay, G. et al., 2003. Nasopharyngeal carriage of *Streptococcus pneumoniae* by adults and children in community and family settings. *Clinical Infectious Diseases*, 38(5), pp.632-639.

Regev-Yochay, G. et al., 2004. Association between carriage of *Streptococcus pneumoniae* and *Staphylococcus aureus* in children. *JAMA*, 292(6), pp.716-720.

Regev-Yochay, G. et al., 2006. Interference between *Streptococcus pneumoniae* and *Staphylococcus aureus*: in vitro hydrogen peroxide-mediated killing by *Streptococcus pneumoniae*. *Journal of Bacteriology*, 188(13), pp.4996-5001.

Reyes, D. et al., 2011. Coordinated regulation by AgrA, SarA, and SarR to control *agr* expression in *Staphylococcus aureus*. *Journal of Bacteriology*, 193(21), pp.6020-6031.

Rohde, H. et al., 2007. Polysaccharide intercellular adhesin or protein factors in biofilm accumulation of *Staphylococcus epidermidis* and *Staphylococcus aureus* isolated from prosthetic hip and knee joint infections. *Biomaterials*, 28(9), pp.1711-1720.

Ronda, C. et al., 1987. Biological role of the pneumococcal amidase. *European Journal of Biochemistry*, 164(3), pp.621-624.

Rooijakkers, S.H.M. et al., 2005. Immune evasion by a staphylococcal complement inhibitor that acts on C3 convertases. *Nature Immunology*, 6(9), pp.920-927.

Rosenow, C. et al., 1997. Contribution of novel choline-binding proteins to adherence, colonization and immunogenicity of *Streptococcus pneumoniae*. *Molecular Microbiology*, 25(5), pp.819-829.

Rosenthal, V.D. et al., 2006. Device-associated nosocomial infections in 55 intensive care units of 8 developing countries. *Annals of Internal Medicine*, 145(8), pp.582-591.

Rubins, J.B. et al., 1995. Dual function of pneumolysin in the early pathogenesis of murine pneumococcal pneumonia. *The Journal of Clinical Investigation*, 95(1), pp.142-150.

Rubins, J.B. et al., 1996. Distinct roles for pneumolysin's cytotoxic and complement activities in the pathogenesis of pneumococcal pneumonia. *American Journal of Respiratory and Critical Care Medicine*, 153(4), pp.1339-1346.

Rudkin, J. et al., 2012. Methicillin resistance reduces the virulence of healthcare-associated methicillin-resistant *Staphylococcus aureus* by interfering with the *agr* quorum sensing system. *The Journal of Infectious Diseases*, 205(5), pp.798-806.

Rudkin, J. et al., 2017. Bacterial toxins: Offensive, defensive, or something else altogether? *PLOS Pathogens*, 13(9), e1006452.

Sadykov, M.R. et al., 2010. Tricarboxylic acid cycle-dependent synthesis of *Staphylococcus aureus* type 5 and 8 capsular polysaccharides. *Journal of Bacteriology*, 192(5), pp.1459-1462.

Saiki, K., Mogi, T. & Anraku, Y., 1992. Heme O biosynthesis in *Escherichia coli*: the *cyoE* gene in the cytochrome *bo* operon encodes a protoheme IX farnesyltransferase. *Biochemical and Biophysical Research Communications*, 189(3), pp.1491-1497.

Saiki, K. et al., 1993. *In vitro* heme O synthesis by the *cyoE* gene product from *Escherichia coli*. *The Journal of Biological Chemistry*, 268(35), pp.26041-26045.

Salipante, S.J. et al., 2015. Large-scale genomic sequencing of extraintestinal pathogenic *Escherichia coli* strains. *Genome Research*, 25(1), pp.119-128.

Salyers, A.A. & Whitt, D.D., 2002. *Bacterial Pathogenesis: A molecular approach*. 2nd ed. Washington D.C.: ASM Press.

Saravolatz, L.D. et al., 1982. Methicillin-resistant *Staphylococcus aureus*: epidemiologic observations during a community-acquired outbreak. *Annals of Internal Medicine*, 96(1), pp.11-16.

Schaetzle, O., Barrière, F. & Baronian, K., 2008. Bacteria and yeasts as catalysts in microbial fuel cells: electron transfer from micro-organisms to electrodes for green electricity. *Energy and Environmental Science*, 1, pp.607-620.

Schaffer, A.C. et al., 2006. Immunization with *Staphylococcus aureus* clumping factor B, a major determinant in nasal carriage, reduces nasal colonization in a murine model. *Infection and Immunity*, 74(4), pp.2145-2153.

Schubert, B. et al., 2019. Genome-wide discovery of epistatic loci affecting antibiotic resistance in *Neisseria gonorrhoeae* using evolutionary couplings. *Nature Microbiology*, 4, pp.328-338.

Shak, J.R., Vidal, J.E. & Klugman, K.P., 2013. Influence of bacterial interactions on pneumococcal colonisation of the nasopharynx. *Trends in Microbiology*, 21(3), pp.129-135.

Sheppard, S.K. et al., 2013. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter*. *PNAS*, 110(29), pp.11923-11927.

Shoma, S. et al., 2008. Critical involvement of pneumolysin in production of interleukin-1α and caspase-1-dependent cytokines in infection with *Streptococcus pneumoniae in vitro*: a novel function of pneumolysin in caspase-1 activation. *Infection and Immunity*, 76(4), pp.1547-1557.

Sibila, O. et al., 2016. Multidrug-resistant pathogens in patients with pneumonia coming from the community. *Current Opinion in Pulmonary Medicine*, 22(3), pp.219-226.

Siboo, I.R. et al., 2001. Clumping factor A mediates binding of *Staphylococcus aureus* to human platelets. *Infection and Immunity*, 69(5), pp.3120-3127.

Simpkin, V.L. et al., 2017. Incentivising innovation in antibiotic drug discovery and development: progress, challenges and next steps. *The Journal of Antibiotics*, 70, pp.1087-1096.

Skovbjerg, S. et al., 2017. Intact pneumococci trigger transcription of interferon-related genes in human monocytes, while fragmented, autolysed bacteria subvert this response. *Infection and Immunity*, 85(5), e00960-16.

Slager, J., Aprianto, R. & Veening, J.W., 2018. Deep genome annotation of the opportunistic human pathogen *Streptococcus pneumoniae* D39. *Nucleic Acids Research*, 46(19), pp.9971-9989.

Slatkin, M., 2008. Linkage disequilibrium – understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6), pp.477-485.

Somerville, G.A. et al., 2002. *Staphylococcus aureus* aconitase activation unexpectedly inhibits post-exponential-phase growth and enhances stationary-phase survival. *Infection and Immunity*, 70(11), pp.6373-6382.

Somerville, G.A. et al., 2003a. Correlation of acetate catabolism and growth yield in *Staphylococcus aureus*: implications for host-pathogen interactions. *Infection and Immunity*, 71(8), pp.4724-4732.

Somerville, G.A. et al., 2003b. Synthesis and deformylation of *Staphylococcus aureus* δ-toxin are linked to tricarboxylic acid cycle activity. *Journal of Bacteriology*, 185(22), pp.6686-6694.

Somerville, G.A. & Proctor, R.A., 2013. Cultivation conditions and the diffusion of oxygen into culture media: the rationale for the flask-to-medium ratio in microbiology. *BMC Microbiology*, 13:9.

Spaan, A.N. et al., 2013. The staphylococcal toxin Panton-Valentine leukocidin targets human C5a receptors. *Cell Host and Microbe*, 13(5), pp.584-594.

Spratt, B.G. & Greenwood, B.M., 2000. Prevention of pneumococcal disease by vaccination: does serotype replacement matter? *The Lancet*, 356(9237), pp.1210-1211.

Steinmoen, H., Teigen, A. & Håvarstein, L.S., 2003. Competence-induced cells of *Streptococcus pneumoniae* lyse competence-deficient cells of the same strain during cocultivation. *Journal of Bacteriology*, 185(24), pp.7176-7183.

Stevens, E. et al., 2017. Cytolytic toxin production by *Staphylococcus aureus* is dependent upon the activity of the protoheme IX farnesyltransferase. *Scientific Reports*, 7:13744.

Sved, J.A. & Hill, W.G., 2018. One hundred years of linkage disequilibrium. *Genetics*, 209(3), pp.629-636.

Titgemeyer, F. et al., 1994. Evolutionary relationships between sugar kinases and transcriptional repressors in bacteria. *Microbiology*, 140(Pt. 9), pp.2349-2354.

Thakker, M. et al., 1998. *Staphylococcus aureus* serotype 5 capsular polysaccharide is antiphagocytic and enhances bacterial virulence in a murine bacteremia model. *Infection and Immunity*, 66(11), pp.5183-5189.

Thet, N.T. et al., 2013. Visible, colorimetric dissemination between pathogenic strains of *Staphylococcus aureus* and *Pseudomonas aeruginosa* using fluorescent dye containing lipid vesicles. *Biosensors and Bioelectronics*, 41, pp.538-543.

Tsuchiya S. et al., 1980. Establishment and characterization of a human acute monocytic leukemia cell line (THP- 1). *International Journal of Cancer*, 26(2), pp.171-176.

Tunjungputri, R.N. et al., 2017. Phage-derived protein induces increased platelet activation and is associated with mortality in patients with invasive pneumococcal disease. *mBio*, 8(1), e01984-16.

Vandenesch, F., Kornblum, J. & Novick, R.P., 1991. A temporal signal, independent of *agr*, is required for *hla* but not *spa* transcription in *Staphylococcus aureus*. *Journal of Bacteriology*, 173(20), pp.6313-6320.

Vandenesch, F. et al., 2003. Community-acquired methicillin-resistant *Staphylococcus aureus* carrying Panton-Valentine leukocidin genes: Worldwide emergence. *Emerging Infectious Diseases*, 9(8), pp.978-984.

Vandenesch, F., Lina, G. & Henry, T., 2012. *Staphylococcus aureus* hemolysins, bi-component leukocidins, and cytolytic peptides: a redundant arsenal of membrane-damaging virulence factors? *Frontiers in Cellular and Infection Microbiology*, 2:12.

van den Oord, E.J.C.G., 2007. Controlling false discoveries in genetic studies. *American Journal of Medical Genetics Part B*, 147B(5), pp.637-644.

van der Poll, T. & Opal, S.M., 2009. Pathogenesis, treatment and prevention of pneumococcal pneumonia. *The Lancet*, 374(9700), pp.1543-1556.

van der Veerdonk, F.L. et al., 2011. Inflammasome activation and IL-1β and IL-18 processing during infection. *Trends in Immunology*, 32(3), pp.110-116.

van Hemert, S. et al., 2010. Identification of *Lactobacillus plantarum* genes modulating the cytokine response of human peripheral blood mononuclear cells. *BMC Microbiology*, 10:293.

Vaudaux, P.E., et al., 1995. Use of adhesion-defective mutants of *Staphylococcus aureus* to define the role of specific plasma proteins in promoting bacterial adhesion to canine arteriovenous shunts. *Infection and Immunity*, 63(2), pp.585-590.

Verbrugh, H.A. et al., 1982. Opsonisation of encapsulated *Staphylococcus aureus*: the role of specific antibody and complement. *The Journal of Immunology*, 129(4), pp.1681-1687.

Vergison, A., 2008. Microbiology of otitis media: a moving target. *Vaccine*, 26(7), pp.G5-G10.

Viney, M., 2014. The failure of genomics in biology. *Trends in parasitology*, 30(7), pp.319-321.

von Eiff, C. et al., 2001. Nasal carriage as a source of *Staphylococcus aureus* bacteraemia. *The New England Journal of Medicine*, 344(1), pp.11-16.

Voyich, J.M. et al., 2006. Is Panton-Valentine leukocidin the major virulence determinant in community-associated methicillin-resistant *Staphylococcus aureus* disease? *The Journal of Infectious Diseases*, 194(12), pp.1761-1770.

Vuong, C. et al., 2000. Impact of the *agr* quorum-sensing system on adherence to polystyrene in *S. aureus*. *The Journal of Infectious Diseases*, 182(6), pp.1688-1693.

Walsh, E.J. et al., 2004. Clumping factor B, a fibrinogen-binding MSCRAMM (microbial surface components recognising adhesive matrix molecules) adhesin of *Staphylococcus aureus*, also binds to the tail region of type I cytokeratin 10. *The Journal of Biological Chemistry*, 279(49), pp.50691-50699.

Wang, B. et al., 2014. Activation and inhibition of the receptor histidine kinase AgrC occurs through opposite helical transduction motions. *Molecular Cell*, 53(6), pp.929-940.

Wartha, F. et al., 2007. Capsule and D-alanylated lipoteichoic acids protect *Streptococcus pneumoniae* against neutrophil extracellular traps. *Cellular Microbiology*, 9(5), pp.1162-1171.

Watson, D.A. & Musher, D.M., 1990. Interruption of capsule production in *Streptococcus pneumoniae* serotype 3 by insertion of transposon Tn*916*. *Infection and Immunity*, 58(9), pp.3135-3138.

Wellmer, A. et a., 2002. Decreased virulence of a pneumolysin-deficient strain of *Streptococcus pneumoniae* in murine meningitis. *Infection and Immunity*, 70(11), pp.6504-6508.

Wertheim, H.F.L. et al., 2005. The role of nasal carriage in *Staphylococcus aureus* infections. *The Lancet Infectious Diseases*, 5(12), pp.751-762.

Whalan, R.H. et al., 2005. PiuA and PiaA, iron uptake lipoproteins of *Streptococcus pneumoniae*, elicit serotype independent antibody responses following human pneumococcal septicaemia. *FEMS Immunology and Medical Microbiology*, 43(1), pp.73-80.

Witzenrath, M. et al., 2011. The NLRP3 inflammasome is differentially activated by pneumolysin variants and contributes to host defense in pneumococcal pneumonia. *The Journal of Immunology*, 187(1), pp.434-440.

World Health Organisation & United Nations International Children's Emergency Fund, 2013. Ending preventable child deaths from pneumonia and diarrhoea by 2025: The integrated global action plan for pneumonia and diarrhoea (GAPPD). [pdf] France: WHO Press. Available at: <http://apps.who.int/iris/bitstream/10665/79200/1/9789241505239_eng.pdf?ua=1> [Accessed 20th June 2016].

Xu, G. et al., 2011. Three *Streptococcus pneumoniae* sialidases: Three different products. *Journal of the American Chemical Society*, 133(6), pp.1718-1721.

Yang, Y. et al., 2019. Recent advances in the mechanisms of NLRP3 inflammasome activation and its inhibitors. *Cell Death and Disease*, 10:128.

Yokoyama, M. et al., 2018. Epistasis analysis uncovers hidden antibiotic resistance-associated fitness costs hampering the evolution of MRSA. *Genome Biology*, 19:94.

Young, B.C. et al., 2012. Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease. *PNAS*, 109(12), pp.4550-4555.

Zafar, M.A. et al., 2017. Host-to-host transmission of *Streptococcus pneumoniae* is driven by its inflammatory toxin, pneumolysin. *Cell Host and Microbe*, 21(1), pp.73-83.

Zhang, L. et al., 2002. Transmembrane topology of AgrB, the protein involved in the post-translational modification of AgrD in *Staphylococcus aureus*. *The Journal of Biological Chemistry*, 277(38), pp.34736-34742.

Zhang, X., Bailey, S.D. & Lupien, M., 2014. Laying a solid foundation for Manhattan – 'setting the functional basis for the post-GWAS era'. *Trends in Genetics*, 30(4), pp.140-149.

Zhu, Y. et al., 2009. Tricarboxylic acid cycle-dependent attenuation of *Staphylococcus aureus* in vivo virulence by selective inhibition of amino acid transport. *Infection and Immunity*, 77(10), pp.4256-4264.